NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS

JAMES HAMILTON VERNER, B.Sc., M.Sc.

A Thesis presented for the

Degree of Doctor of Philosophy

of the University of Edinburgh

in the Faculty of Science

October, 1969.

PREFACE

In October 1966, the author commenced study as an Internal Student for the degree of Ph.D. in Numerical Analysis under the supervision of Dr. M.J.M. Bernal at the University of London Institute of Computer Science. Subsequently the research which is the subject of this thesis was started at the suggestion of Dr. G.J. Cooper. Dr. Cooper continued supervision of this research in the Department of Computer Science at the University of Edinburgh where the author was admitted as a student under the Ph.D. Ordnance in October 1967.

The author acknowledges with deep gratitude the academic and moral encouragement given by his supervisors, especially Dr. Cooper whose discussions resulted in a stimulating research project. The author acknowledges the financial support for this research provided in the form of a postgraduate scholarship from the National Research Council of Canada. Thanks are due to Mrs. Ray Chester for her technical reproduction of this thesis.

SUMMARY

Recent investigations by Butcher have co-ordinated several aspects of the theory of, and techniques for the derivation of, numerical methods for ordinary differential equations, and further have yielded several significant classes of new methods. It appears that the fruitfulness of this research is, in part, a result of the generality with which this author has approached his subject. Using the same sort of general approach, several applications and generalizations in this field of research are developed as the subject of this thesis.

Initially a system of explicit differential equations of arbitrary orders for initial value problems is considered, and a suitable notation is used to express a general class of numerical methods for such problems. Stability and consistency are defined for methods of this class, and it is shown that a method satisfying these conditions is convergent.

Several important existence theorems are based on the construction of certain numerical solutions. Indeed several authors have proved existence of solutions to first order systems using particular subclasses of methods. Here Peano's existence theorem for first order systems is derived, using a general class of methods.

Several new results are then proved for particular subclasses of the general class of methods. An examination of a particular group of explicit Runge-Kutta methods of order six (jointly with G.J. Cooper) has yielded a corresponding group of methods of order

eight. Certain implicit methods for a system of equations of arbitrary orders are shown to be of maximum order; this is a generalization of a result proved by Butcher for first order systems. A similar result is given for a more general class of methods (to be referred to as "quadrature" methods in this thesis) in which weight functions may be used to compensate for difficult behaviour. A recursion is developed to generate stable hybrid methods for a system of differential equations of arbitrary orders from corresponding methods for a first order system. A comparison shows that, in general, such new methods do not improve on the technique of reducing a system of equations of arbitrary orders to one of first order, and then using a method for a first order system.

Quadrature methods are applied to find numerical solutions for certain systems of implicit differential equations. This approach yields a useful reformulation of an existence theorem for implicit problems. Further, it appears that the numerical results obtained for problems with certain difficult behaviour may not be obtained by other known techniques.

Finally (in less generality) consideration is given to the error in numerical solutions for explicit differential equations. An efficient algorithm for the adjustment of steplength to control error is developed for certain classes of single step methods, and several examples are given for explicit methods. Finally, an easily computed error bound for a certain class of problems is given for explicit single step methods.

DECLARATION

I declare that this thesis has been composed by myself, and that the original results referred to in the summary are my own contribution.

September, 1969. James H. Verner.

C O N T E N T S

C O N T E N T S (Contd.)

C O N T E N T S (Contd.)

C O N T E N T S (Contd.)

INTRODUCTION

The first method for the numerical solution of a differential equation was developed by Euler in 1768. Fifty years later Cauchy used this method in proving the existence of solutions for a certain class of differential equations. Indeed the use of numerical solutions in existence theorems has stimulated research in both directions. (The history and development of existence theorems for ordinary differential equations is discussed by Sarafyan [37] .) The basic existence theorems for initial value problems were developed during the nineteenth century, while more recent research has been concerned with the qualitative behaviour of solutions.

Interest in numerical solutions was established with the development of a multi-step method by Adams* [3] in 1883, and of single step methods by Runge [35, 36] and Kutta [27], at the turn of the century. Since then, many new methods have been proposed. Although some of these methods are suitable for special problems, there are, as yet, no entirely adequate general purpose methods. Indeed even now the Adams-Bashforth and classical Runge-Kutta methods form the basis of standard routines on many digital computers for obtaining numerical solutions of differential equations.

---

* This is traditionally referred to as the Adams-Bashforth method. It was developed for a particular problem investigated by the latter.

Recent investigations in this field have been more fruitful. Dahlquist [20] was the first to treat a class of (multi-step) methods as such, and defined conditions under which a multi-step method was convergent. Then there appeared a number of monographs, which included analyses of methods for ordinary differential equations. Notable among these are Henrici's books [23, 24] which give an excellent account of the results known for initial value problems at the time. Subsequent research has yielded several new classes of methods including implicit single step methods and hybrid methods. Butcher's contributions [6-13] are fundamental. Indeed the generality of his approach has co-ordinated much of the analysis and derivation of numerical methods, and further led to the development of several new subclasses of methods.

It appears that the fruitfulness of this research is partially due to the generality of approach taken by the various authors. Indeed, many results in this thesis are derived by considering classes of methods as such, and could not have been otherwise developed easily.

The problems to be considered are classified at the beginning of Chapter I. This chapter further includes statements of basic existence and uniqueness theorems for initial value problems. Using a general class of numerical methods, an existence theorem for first order systems is proved at the end of Chapter II.

Chapter II commences with a notation for a general class of numerical methods. Conditions under which such methods yield unique numerical solutions are given. Stability and consistency

are defined for these methods, and it is shown that a stable, consistent method is convergent. The use of a convergent method is as fundamental to solving a differential problem numerically as is the requirement that there exists a solution to the differential problem.

Each of the next four chapters is concerned with a particular subclass of the general methods. We begin with explicit Runge-Kutta methods, for which parameter constraints are derived to determine order. A recursion is given from which parameters for a method (of the same order) for a system of differential equations of arbitrary orders may be generated from those of a method for a first order system. Then the parameters for a group of methods of order six, given by Butcher [7], are examined in detail, and the existence of an associated subclass of methods of arbitrary even orders is proposed. The method of order four for the class is given, and a group of methods of order eight is derived. It appears that methods of higher orders of this subclass do not exist, although no proof is given.

In a similar way, parameter constraints to determine the order of implicit Runge-Kutta methods are given in the next chapter. Butcher [8, 9] has shown that certain subclasses of implicit methods for first order systems are of maximum order. This result is proved here for associated implicit methods for systems of equations of arbitrary orders. Cooper [16] has shown that parameters for some of these methods may be generated by the recursion given for explicit methods. In a numerical example,

the explicit methods derived in Chapter III are compared with closely related implicit methods.

Quadrature methods form a more general class of implicit methods in which weight functions may be used to compensate for difficult behaviour (Cooper [17]). Again parameter constraints to determine the order of a method are expressed in an appropriate form. It is then shown that certain of the methods are of maximum order.

Hybrid methods form a recently derived class in which the virtue of the stability* of single step methods of maximum order is combined with that of a minimum of function evaluations required for multi-step methods. Two recursions are derived which are similar in purpose, but more general in nature than that given for explicit Runge-Kutta methods. It is indicated that the first of these recursions is inferior; indeed, for problems in which fourth or higher order derivatives are included, the corresponding methods generated are unstable. Fortunately, methods generated by the second recursion are completely stable, and further have the same stability characteristics as the associated methods for first order systems. In general, it appears that methods derived by either recursion do not improve on the classical technique of reducing a system of equations of arbitrary orders to one of first order, and using a method for

---

* Stability of a method is defined later; all consistent single step methods are stable.

first order systems. However, these methods are often more efficient in the solution of problems in which one or more lower order derivatives do not appear explicitly in the system.

It appears that only a few investigators have attempted the numerical solution of a system of implicit differential equations - that is a system in which the derivatives of highest order do not occur explicitly. In Chapter VII an apparently useful application of quadrature methods is made to find numerical solutions for implicit differential equations directly. As a result of this investigation a basic existence theorem for implicit differential equations is reformulated to include a larger class of problems. Convergence of the methods is proved, and several numerical examples indicate that certain methods (based on quadratures) may be of maximum order.

Recently special techniques have been developed for obtaining numerical solutions to problems in which a rapid variation of the solution or a derivative occurs within a sub-interval of the interval on which the solution is required. In such problems an attempt is made to control error by appropriately adjusting the steplength whenever a significant change occurs in a local error estimate. It appears that multi-step methods are, in general, unsuitable in any sub-interval where the variation is extreme. Indeed, the adjustment itself is often difficult, although this difficulty may be avoided by using a method such as that proposed by Nordsieck [34]. Further, in multi-step methods, approximations to the higher-ordered

derivatives are available only at previously considered points of the integration interval; for extreme problems such approximations may not be adequate. Both disadvantages are avoided by single step methods, but for such methods the error estimates available require considerable computation. In Chapter VIII a relative error estimate requiring little computation is proposed, and from the numerical examples given it appears that its use is justified.

To complete the analysis of a numerical solution a bound for the error is needed. In general, realistic error bounds are not available. In deriving error bounds, there are fundamental difficulties which are not avoided here. However, in Chapter IX, a very restricted class of problems is considered. For this class, an easily computed, but pessimistic, error bound is derived for explicit Runge-Kutta methods. With a classification of other problems, future research may yield bounds of a similar nature. Certainly, further research in this direction is necessary.

Results for the numerical examples in this thesis were obtained from the Edinburgh Regional Computing Centre KDF9. The programs were written in Atlas Autocode using single length (39 significant binary digits) or double length (78 significant binary digits) floating point arithmetic. In tabulating results, numbers in brackets represent exponents to base 10.

CHAPTER I

ORDINARY DIFFERENTIAL EQUATIONS

1. Introduction

Prior to selecting a numerical method and determining the numerical solution of a differential problem, it is necessary to establish that there exists a solution to the problem. Further, if the solution is unique, then a numerical method defining a unique approximation for a fixed steplength defines an approximation to this solution. It may not, a priori, be possible to determine that a unique solution exists. Indeed, if a Lipschitz condition (defined below) is not satisfied, an initial value problem may have more than one solution. Further, Sarafyan [37] discusses the possibility of obtaining numerical approximations to the two "extremal" solutions of such a problem. For such problems, special consideration is necessary if numerical solutions are to be correctly interpreted; such consideration is beyond the scope of this thesis.

It is convenient to begin with a classification of differential equations, and the problems derived therefrom. Then basic existence theorems for the classes of problems to be considered may be stated, and further, numerical methods are more easily examined for each class of problems individually.

2. Classification

Consider a real linear n-dimensional space, R_n, of vectors $\underline{y} \equiv (y_1, y_2, \ldots, y_n)$. For spaces of finite dimension, all norms are equivalent. Here, several results are more easily derived using the uniform norm defined by

$$\|\underline{y}\| = \max_{1 \le i \le n} |y_i| .$$

Now a general system of ordinary differential equations of arbitrary orders is defined by

$$(1.1) \qquad F_r(t; \underline{y}(t), \underline{z}(t)) = 0 , \qquad r = 1(1)q,$$

where

$$\underline{y}(t) \equiv \left\{ y_\rho^{(n_\rho - m)}(t) \right\} =$$

$$= (y_1(t), \ldots, y_1^{(n_1-1)}(t); \ldots; y_q(t), \ldots, y_q^{(n_q-1)}(t))$$

is a point of R_N with $N = n_1 + \ldots + n_q$, and

$$\underline{z}(t) \equiv \left\{ y_\rho^{(n_\rho)}(t) \right\} = (y_1^{(n_1)}(t), \ldots, y_q^{(n_q)}(t))$$

is a point of R_q. If (1.1) has the form

$$(1.2) \qquad y_r^{(n_r)}(t) = f_r(t; \underline{y}(t)), \qquad r = 1(1)q,$$

then it is a system of explicit differential equations; otherwise, it is a system of implicit differential equations. Further, (1.2) is a system of linear differential equations if

$$(1.3)\qquad y_r^{(n_r)}(t) = f_r(t; y(t)) \equiv \sum_{\rho=1}^{q} \sum_{i=0}^{n_\rho - 1} a_{r\rho i}(t)\, y_\rho^{(i)}(t) + a_r(t) ,$$

$$r = 1(1)q .$$

Unless otherwise implied, we shall mean a system in the form (1.2) when referring to a system of differential equations.

A differential problem is defined by a system of differential equations together with values of the components of $\underline{y}(t)$ and $\underline{z}(t)$ for one or more values of t. Here, we consider only initial value problems, for which these values are given for only one value of $t = x$.

The orders of a system of differential equations refer to the set of orders of the highest-ordered derivatives of the system, $(n_1, \ldots, n_q)$. This concept is different and distinct from that of the order of a numerical method defined in Chapter II.

3. Existence and Uniqueness

Theorems implying the existence and uniqueness of solutions are to be found in most monographs on the theory of ordinary differential equations. Murray and Miller [32] have collected a number of these results together, and statements for several of these are included here. Some existence theorems with less restrictive hypotheses are available for special classes of problems; for example Abian and Brown [1, 2] prove the existence of unique solutions for certain implicit differential equations. However, as several results for numerical methods require the same hypotheses as the basic existence theorems, it will be assumed that one of these is valid for any problem considered.

Theorem (1.1): Consider the hypotheses:

H1: $f_r(t; y_1, \ldots, y_q)$, $r = 1(1)q$,

are q real valued functions of the $(q+1)$ real variables, defined and continuous on an open region D of R_{q+1}.

H2: (Lipschitz condition). There exists a constant L such that for every pair of points $(t; \underline{y}^*)$ and $(t; \underline{y})$ in D, we have

$$\left| f_r(t; \underline{y}^*) - f_r(t; \underline{y}) \right| \leq L \left\| \underline{y}^* - \underline{y} \right\| , \qquad r = 1(1)q .$$

Then under hypothesis H1, for every point $(x; \underline{y}_o)$ of D, we can find $b > 0$ and q functions $\phi_1(t), \ldots, \phi_q(t)$, which have continuous first derivatives in the neighbourhood

N_1 : $|t - x| \leq b$ such that

$$\frac{d\phi_r(t)}{dt} = f_r(t; \phi_1(t), \ldots, \phi_q(t)), \qquad r = 1(1)q,$$

in this neighbourhood, and

$$\phi_r(x) = y_{ro}, \qquad r = 1(1)q.$$

Furthermore, H2 implies this is the only set of functions having these properties.

The proof (see Murray and Miller [32, p. 13]) proceeds by the construction of a sequence of bounded equicontinuous functions approximating the exact solution. By Arzela's theorem (see, for example, Kolmogorov and Fomin [26, p. 54]), it follows that the sequence converges, and existence is established by showing that the limit of the sequence has a derivative, and further, satisfies the differential equation. Uniqueness is proved by writing the differential equation equivalently as an integral equation, and invoking the Lipschitz condition to show that any two solutions are identical. The results are extended to a first order system of implicit differential equations in the next theorem, and then to a system of arbitrary orders.

Theorem (1.2): Consider the hypotheses:

H1: $F_r(t; y_1, \ldots, y_q, y_1', \ldots, y_q')$, $r = 1(1)q,$

are q functions of the (2q+1) real variables, defined and

continuous on an open region D of R_{2q+1}.

H2: $\frac{\partial F_r}{\partial y_\rho'}$ exist and are continuous on D for $r, \rho = 1(1)q$.

H3: There exists a point $(x; \underline{Y}_o, \underline{z}_o)$ in D such that

$$F_r(x; \underline{Y}_o, \underline{z}_o) = 0 , \qquad r = 1(1)q,$$

and the Jacobian

$$J = \frac{\partial(F_1, \ldots, F_q)}{\partial(y_1', \ldots, y_q')} \neq \begin{vmatrix} \frac{\partial F_1}{\partial y_1'} & \cdots & \frac{\partial F_1}{\partial y_q'} \\ \vdots & & \vdots \\ \frac{\partial F_q}{\partial y_1'} & \cdots & \frac{\partial F_q}{\partial y_q'} \end{vmatrix}$$

is not zero at this point.

H4: $\frac{\partial F_r}{\partial y_\rho}$ exist and are continuous on D for $r, \rho = 1(1)q$.

Then under hypotheses H1, H2, and H3, there exists $b' > 0$ such that for every set of $(q+1)$ values $x^*, y_1^*, \ldots, y_q^*$ with $|y_r^* - y_{ro}| < b'$ and $|x^* - x| < b'$, there exists $b^* > 0$ $(b^* \leq b')$ and q functions $\phi_1(t), \ldots, \phi_q(t)$, with

$$y_r^* = \phi_r(x^*), \qquad r = 1(1)q ,$$

which have continuous first derivatives in the neighbourhood,

N_2: $|t - x^*| \leq b^*$ such that

$$F_r(t; \phi_1, \ldots, \phi_q, \frac{d\phi_1}{dt}, \ldots, \frac{d\phi_q}{dt}) \equiv 0 \qquad r = 1(1)q.$$

Furthermore, hypothesis H4 implies these functions are unique.

Murray and Miller [32, p. 28] prove the theorem by invoking the implicit function theorem to show that the system of implicit equations may be treated as a system of explicit equations; then existence follows using Theorem (1.1). Using H4, a Lipschitz condition is established, thereby proving uniqueness.

The basic existence theorem for solutions to the system (1.1) is now stated (Murray and Miller [32, p. 32]).

Theorem (1.3): Consider the hypotheses:

H1: $F_r(t; \underline{y}(t), \underline{z}(t)) = 0 \quad , \qquad r = 1(1)q,$

are q real valued functions of the (N+q+1) real variables

$$t, y_1, \ldots, y_1^{(n_1-1)}, \ldots, y_q, \ldots, y_q^{(n_q-1)}, y_1^{(n_1)}, \ldots, y_q^{(n_q)},$$

defined and continuous on an open region D of R_{N+q+1}.

H2: $\dfrac{\partial F_r}{\partial y_\rho^{(n_\rho)}}$ exists and is continuous on D for $r, \rho = 1(1)q.$

H3: There exists a point $(x; \underline{y}_o, \underline{z}_o)$ in D such that

$$F_r(x; \underline{y}_o, \underline{z}_o) = 0, \qquad r = 1(1)q,$$

and further the Jacobian

$$J = \frac{\partial(F_1, \ldots, F_q)}{\partial(y_1^{(n_1)}, \ldots, y_q^{(n_q)})}$$

is not zero at this point.

H4: $\dfrac{\partial F_r}{\partial y^{(n_\rho - \nu)}}$ exists and is continuous on D for

$r, \rho = 1(1)q, \quad \nu = 1(1)n_\rho$.

Under the hypotheses, the given system is equivalent to a first order system of N equations, and further, the hypotheses of Theorem (1.2) are satisfied.

Equivalence follows by assuming, for example, $n_1 > 1$, setting

$$\omega_1 = y_1, \quad \omega_2 = y_1', \ \ldots, \ \omega_{n_1} = y_1^{(n_1-1)},$$

and supplementing the system of q equations with

$$\omega_\nu' = \omega_{\nu+1}, \qquad \nu = 1(1)n_1-1 .$$

In Chapter VII, Theorem (1.3) is extended to include a larger class of problems. As this extension arose from associated numerical methods, it appears more suitable to include it with that derivation.

Most numerical methods are derived for first order systems, and for such methods, a system of equations of arbitrary orders must be reduced to a first order system by the technique above

(or otherwise). However, methods can be specially constructed for higher order systems. For example, de Vogeleare [41] derived a method for second order systems, in which the first derivative did not appear, and Cooper [16, 17] derived methods for systems of arbitrary orders.

All of the above results are local, in the sense that the problem itself dictates the interval over which existence is ensured. In general, a solution cannot be continued indefinitely; additional conditions are needed (see Nemytskii and Stepanov [33, p. 8]). A special case follows.

Theorem (1.4): For the system (1.3), let $a_{r\rho i}(t)$ be defined and continuous on some closed interval $I = [a, b]$ of the real variable t. Then for any point $x \in I$, and any vector $\underline{y}_o$, there exists a unique solution of (1.3), $\underline{y}(t)$, for $t \in I$ such that

$$\underline{y}(x) = \underline{y}_o.$$

To prove the theorem, Murray and Miller [32, p. 112] reduce the system to a first order system as in Theorem (1.3). Continuity of the coefficients is used to show the existence of a Lipschitz condition for all $t \in I$. Existence and uniqueness is proved successively in a finite number of sub-intervals of equal length (except possibly at the end points of I) which cover I, and linearity is used to show that the solution is uniquely continued from one sub-interval to the next.

4. Other Existence Theorems

In general, proofs of existence theorems are based on showing that a sequence of functions converges uniformly to the solution of a given system of differential equations. Indeed the classical proof of Cauchy shows that a sequence of numerical approximations converges to a unique solution; Picard (see Murray and Miller [32, p. 59]), on the other hand, uses a sequence of analytical approximations to the same end. It appears natural to expect that such a sequence, truncated after a large number of terms, provides a "good" approximation to the exact solution. It appears just as natural to assume that an existence theorem may be proved using an appropriate sequence of approximations, for example, such as is generated by a numerical method in decreasing the steplength. Such proofs are given by Hull and Luxemburg [25] for multi-step methods, and by Sarafyan [37] for certain single step methods for first order systems. A similar result for a more general class of methods is given in the next chapter after certain concepts are defined.

CHAPTER II

NUMERICAL METHODS FOR EXPLICIT DIFFERENTIAL EQUATIONS

1. Introduction

Henceforth, we assume that a given differential problem has a unique solution, and, in particular, satisfies a Lipschitz condition*. A numerical method specifies approximate values for the solution of the problem at a discrete set of points of some interval. An approximate solution may then be defined throughout the interval, for example, by interpolation. Here, conditions under which the approximation is "close to" the actual solution are examined; indeed, such conditions must be satisfied by a method if the numerical solution is to be acceptable.

Butcher [12] expresses numerical methods for the solution of first order differential equations in a notation which includes both single step and multi-step methods, as well as methods of a more general type. He defines stability and consistency for these methods, and shows that a stable, consistent method is convergent. Cooper [16, 17] develops several classes of methods for a system of equations of arbitrary orders. Here, a representation for more general methods for such a system is developed in the matrix formulation used by Butcher. Stability

---

* In Section 6, an existence theorem is proved, and for the proof, neither of the above assumptions is made.

and consistency are defined, and it is shown that a stable, consistent method is convergent for systems of arbitrary orders. In following chapters, subclasses of these methods are considered. Here, it is convenient to revert to the conventional notation.

2. Notation

For a vector $\underline{y}_o$, we consider an initial value problem in which the differential system has the form (1.2), and for which there exists a unique solution $\underline{y}(t)$ in some interval I containing a point x with

$$\underline{y}(x) = \underline{y}_o .$$

It is assumed that the functions $f_r(t; \underline{\omega})$, $r = 1(1)q$, are continuous and satisfy a Lipschitz condition: there exists a constant L such that

$$\left| f_r(t; \underline{\omega}) - f_r(t; \underline{\omega}') \right| \leq L \left\| \underline{\omega} - \underline{\omega}' \right\| , \qquad r = 1(1)q ,$$

for any two points $(t; \underline{\omega})$ and $(t; \underline{\omega}')$ of R_{N+1} with $t \in I$. Also for some positive integer p, the derivatives

$$(2.1) \qquad y_r^{(n_r+p)}(t) , \qquad r = 1(1)q ,$$

exist and are continuous in I.

A set of abscissae $\{\mu_i \mid i = 1(1)s\}$ will be associated with a method (a method requiring s abscissae is referred to as an s-stage method); it is further convenient to define $\mu_{s+1} = 1$. For given (or previously determined) values $\underline{y}(x)$, a numerical method determines approximations $\hat{\underline{y}}(x')$ to $\underline{y}(x')$ at some point $x' \in I$, $x' \neq x$. The steplength defined by $h = x' - x$ is (usually) a small increment in the variable t, and defines a set of points:
$\{ x_i = x + \mu_i h \mid i = 1(1)s+1 \}$, and an interval $I' = [a', b']$ where

$$a' = \min_i \{ x_i \} , \qquad b' = \max_i \{ x_i \} .$$

As a numerical solution is required in I, x' (and also h) is chosen so that $I' \subseteq I$; further, we shall consider $\bar{x}$ with $\bar{x} - x = mh$ for some positive integer m, and require that $\bar{x} + \mu_i h \in I$, $i = 1(1)s+1$.

It is now convenient to introduce a matrix notation in which a numerical method may be expressed. We define column vectors

$$Y_r^{(\eta)} (t, \underline{\mu}h), \qquad r = 1(1)q, \quad \eta = 0(1)p+n_r ,$$

of R_{s+1} where the i-th component is given by

$$(2.2) \qquad \left[Y_r^{(\eta)} (t, \underline{\mu}h) \right]_i = y_r^{(\eta)} (t + \mu_i h) ,$$

$$r = 1(1)q, \quad \eta = 0(1)p+n_r, \quad i = 1(1)s+1 ,$$

and whenever there is no ambiguity, $\underline{\mu}$ will be omitted.

Certain linear transformations of R_{s+1} to R_{s+1} are defined by the matrix

$$M = \begin{bmatrix} \mu_1 & \cdots & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & \mu_{s+1} \end{bmatrix} ,$$

and matrices*

$$A_r^{[\nu,\tau]} = \begin{bmatrix} \lambda_{r11}^{[\nu,\tau]} & \cdots & \lambda_{r1s}^{[\nu,\tau]} & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \lambda_{rs1}^{[\nu,\tau]} & \cdots & \lambda_{rss}^{[\nu,\tau]} & 0 \\ a_{r1}^{[\nu,\tau]} & \cdots & a_{rs}^{[\nu,\tau]} & 0 \end{bmatrix} ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad \tau = 0(1)\nu ,$$

where the parameters $\left\{ \lambda_{rij}^{[\nu,\tau]}, a_{ri}^{[\nu,\tau]} \right\}$ are defined (implicitly) in the next section. A norm on the matrices is defined by

$$\| M \| = \max_i \left\{ \sum_{j=1}^{s+1} |m_{ij}| \right\} ;$$

this norm is subordinate to the vector norm. For non-linear transformations

$$\bar{f}_r(\underline{x}; \underline{Y}(x,h)) , \qquad r = 1(1)q,$$

---

* The final columns of these matrices need not be chosen zero for (2.4); however as this is always the case for (2.6), this is no restriction on the class of methods considered.

of $R_{(N+1)(s+1)}$ to R_{s+1}, the i-th component is defined by

$$\left[\bar{f}_r\ (\underline{x};\ \underline{Y}(x,h))\right]_i \ = \ f_r\ (x_i\ ;\ y(x_i)), \qquad i = 1(1)s+1,$$

where vectors defined by their transposes are given by

$$\underline{x}^T \ = \ (x_1, \ldots, x_{s+1})$$

and

$$\underline{Y}(x,h)^T = (Y_1(x,h)^T, \ldots, Y_1^{(n_1-1)}(x,h)^T; \ldots;$$
$$Y_q(x,h)^T, \ldots, Y_q^{(n_q-1)}(x,h)^T).$$

3. Numerical Methods

The derivation of a numerical method is motivated, in part, by the desire to obtain an approximation as close as possible to the actual solution. Indeed, the derivation is based on an a priori assumption of the form of the error. Here, it is assumed that the error occurs in the final term of a truncated Taylor series for the approximation. For certain functions $\{f_r\}$, such methods give exact results for a polynomial (solution) of appropriate degree. Other types of approximation are sometimes preferable; for example, Lambert and Shaw [28] derive methods based on rational function approximation which give better results near to a singularity. Certain additional conditions (discussed in Section 5) must be imposed if a numerical method is to be acceptable.

Truncated Taylor expansions for the vectors (2.2) may be written as

$$(2.3)\qquad Y_r^{(n_r-\nu)}(x,\underline{\mu}h) = \sum_{\sigma=0}^{\bar{p}} \frac{h^\sigma}{\sigma!} M^\sigma Y_r^{(n_r-\nu+\sigma)}(x,0) + \frac{h^{\bar{p}+1}}{(\bar{p}+1)!} M^{\bar{p}+1} Y_r^{(n_r-\nu+\bar{p}+1)}(x,\underline{\theta}h)$$

$$r = 1(1)q, \quad \nu = 1(1)n_r ,$$

where $\underline{\theta}$ is a vector independent of h with $|\theta_i| < |\mu_i|$, $i = 1(1)s+1$, and $\bar{p}$ is a non-negative integer. By (2.1) the derivatives in (2.3) are continuous, and thus bounded for $0 < \bar{p} \leq p$. For $|h|$ sufficiently small, the final term of (2.3) is negligible with respect to that for $\sigma = 0$ (whenever $\underline{Y}_0 \neq \underline{0}$); by neglecting this term, (2.3) with $\bar{p} = \nu$ gives the required approximations. Approximating $\underline{Y}(x, 0)$ by a linear transformation of $\underline{Y}(x, \underline{\mu}h)$ leads to more general methods given by

$$(2.4)\qquad \hat{Y}_r^{(n_r-\nu)}(x,h) = \sum_{\sigma=0}^{\nu} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} \hat{Y}_r^{(n_r-\nu+\sigma)}(x,h)$$

$$r = 1(1)q, \quad \nu = 1(1)n_r ,$$

where the approximations for $Y_r^{(n_r)}(x,h)$ are given by

$$(2.4')\qquad \hat{Y}_r^{(n_r)}(x,h) = \bar{f}_r(\underline{x}; \hat{\underline{Y}}(x,h)) , \qquad r = 1(1)q.$$

It will be seen in later chapters that the parameters of the transformations $A_r^{[\nu,\sigma]}$ may be chosen so that the Taylor expansion of (2.4) with $\hat{\underline{Y}}(x,0) = \underline{Y}(x,0)$ agrees with that in (2.3) for $\bar{p} > \nu$. We assume for the present that this is possible.

We define a function $E(h)$ of a real variable to have the property

$$E(h) = O(h^{p+1})$$

if

$$\lim_{h \to 0} \frac{E(h)}{h^{p+1}} = \text{constant.}$$

This leads to the concept of the order of a numerical method.

Definition (2.1): A numerical method defined by (2.4) is of order p if

$$(2.5) \quad \left[\hat{Y}_r^{(n_r-\nu)}(x,h) - Y_r^{(n_r-\nu)}(x,h) \right]_{s+1} = O(h^{p+1}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

It is necessary only that the error of the required approximations $\hat{\underline{y}}(x + h)$ be of order p, for a method to be of order p. The other components of $\hat{\underline{Y}}(x,h)$ are defined only to obtain these approximations, and the order of their error is otherwise immaterial. It will be seen that certain methods of arbitrary order are known, and thus for any problem the maximum order of the error is restricted only by the value of p for which the derivatives (2.1) are continuous.

4. Uniqueness of the Numerical Solution

A particular numerical method from (2.4) is defined by the choice of parameters of the transformations $A_r^{[\nu,\sigma]}$.

Here, conditions are given under which (2.4) uniquely determines the unknown components of $Y_r^{(n_r-\nu)}(x,h)$. If these conditions are satisfied, and if the functions

$$f_r(t; \underline{Y}), \qquad r = 1(1)q,$$

are linear in $\underline{Y}$, (2.4) forms a system of simultaneous linear equations, and may be solved directly. However, in general, the system will be non-linear, and must be solved iteratively.

Now (2.4) has the form

$$\hat{\underline{Y}} = G(\hat{\underline{Y}}) = \left\{ G_r^{[\nu]}(\hat{\underline{Y}}) \right\},$$

and does not define $\hat{\underline{Y}}$ explicitly. However, if G is a contraction mapping, this vector is uniquely determined by the iteration

$$\hat{\underline{Y}}^{(k+1)} = G(\hat{Y}^{(k)}), \qquad k = 0, 1, \ldots .$$

Theorem (2.1): Let $\bar{A}_r^{[\nu,o]}$ be defined by replacing the i-th column of $A_r^{[\nu,o]}$ with zeros whenever the i-th component of $Y_r^{(n_r-\nu)}(x,h)$ is given. If

$$\left\| \bar{A}_r^{[\nu,o]} \right\| < 1, \qquad r = 1(1)q, \quad \nu = 1(1)n_r,$$

then for $|h|$ small enough, the transformation defined by (2.4) is a contraction mapping.

Proof: Let $\underline{V}$ and $\underline{W}$ be any points of $R_{N(s+1)}$ for which components corresponding to given values of $\underline{Y}(x, h)$ are equal (for in using the iteration these components will always be equal to the given values). Then we may write

$$G_r^{[\nu]}(\underline{V}) - G_r^{[\nu]}(\underline{W})$$

$$= \bar{A}_r^{[\nu,0]}\left[V_r^{(n_r-\nu)} - W_r^{(n_r-\nu)}\right] + \sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]}\left[V_r^{(n_r-\nu+\sigma)} - W_r^{(n_r-\nu+\sigma)}\right]$$

$$+ \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]}\left[\bar{f}_r(\underline{x};\underline{V}) - \bar{f}_r(\underline{x};\underline{W})\right] ,$$

$$r = 1(1)q \quad , \qquad \nu = 1(1)n_r .$$

Using the Lipschitz condition this gives

$$\left\|G_r^{[\nu]}(\underline{V}) - G_r^{[\nu]}(\underline{W})\right\|$$

$$\leq \left\{\left\|\bar{A}_r^{[\nu,0]}\right\| + \sum_{\sigma=1}^{\nu-1} \frac{|h|^\sigma}{\sigma!}\|M\|^\sigma\left\|A_r^{[\nu,\sigma]}\right\| + L\frac{|h|^\nu}{\nu!}\|M\|^\nu\left\|A_r^{[\nu,\nu]}\right\|\right\}\left\|\underline{V}-\underline{W}\right\| ,$$

$$r = 1(1)q, \qquad \nu = 1(1)n_r .$$

By the definition of the vector norm, it follows that

$$\left\|G(\underline{V}) - G(\underline{W})\right\| = \max_{r,\nu}\left\{\left\|G_r^{[\nu]}(\underline{V}) - G_r^{[\nu]}(\underline{W})\right\|\right\}$$

$$\leq \beta(h)\left\|\underline{V} - \underline{W}\right\| ,$$

where for $|h|$ sufficiently small, $\beta(h) < 1$ and thus G is a contraction mapping.

A unique solution for (2.4) may exist even if the transformation is not a contraction mapping. However, it will be assumed that there is a contraction mapping, for in this case (2.4) may be rewritten. Indeed, if $\left\|\bar{A}_r^{[\nu,0]}\right\| < 1$, then $(I - \bar{A}_r^{[\nu,0]})^{-1}$ exists, and (2.4) becomes

$$\hat{Y}_r^{(n_r-\nu)}(x,h) = (I - \bar{A}_r^{[\nu,0]})^{-1} \Big[(A_r^{[\nu,0]} - \bar{A}_r^{[\nu,0]}) \hat{Y}_r^{(n_r-\nu)}(x,h)$$

$$+ \sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} \hat{Y}_r^{(n_r-\nu+\sigma)}(x,h)$$

$$+ \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]} \bar{f}_r(\underline{x}; \hat{\underline{Y}}(x,h)) \Big]$$

$$r = 1(1)q, \qquad \nu = 1(1)n_r \ .$$

Further as the first term in square brackets is given, it may be written as a transformation of $\hat{Y}_r^{(n_r-\nu)}(x-h, h)$. Thus with a change of variable, and a modification of the linear transformations, methods to be considered may be written

$$(2.6)\quad \hat{Y}_r^{(n_r-\nu)}(x+h,h) = A_r^{[\nu,0]} \hat{Y}^{(n_r-\nu)}(x,h) + \sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} \hat{Y}_r^{(n_r-\nu+\sigma)}(x+h,h)$$

$$+ \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]} \bar{f}_r(\underline{x}+h\underline{s}; \hat{\underline{Y}}(x+h,h)) \ ,$$

$$r = 1(1)q, \qquad \nu = 1(1)n_r \ ,$$

where now we redefine (2.2) by

$$(2.2')\quad \left[Y_r^{(\eta)}(t, \underline{\mu}h)\right]_i = y_r^{(\eta)}(t + (\mu_i - 1)h) \ ,$$

$$r = 1(1)q, \quad \eta = 0(1)p+n_r, \quad i = 1(1)s+1 \ ,$$

and $\underline{s}$ is a vector with every element equal to unity.

This representation is much more useful than (2.4). The solutions of both are identical; however the rate of convergence of the corresponding iteration for the latter depends (approximately) on the reciprocal of h, and is much faster than that

for (2.4). Further, Butcher [12] has shown for first order systems that (2.6) leads to an elegant consideration of stability*, and therefore convergence*, of a method. The analogous treatment for systems of arbitrary orders is given in the next section.

5. Stability, Consistency and Convergence

A criterion for "local" closeness of a numerical solution to the actual solution motivated the derivation of numerical methods. If a method is to be useful, it is necessary (and sufficient) that the numerical solution be close to the actual solution throughout an interval I. Repetitive application of (2.6) defines approximations $\hat{\underline{Y}}(x+kh,h)$, $k = 1,2,\ldots, m$. Although we proceed to discuss convergence in terms of these tabular functions, some procedure may be defined to interpolate between these values. Indeed, such a procedure is used in an existence theorem later.

Definition (2.2): A method given by (2.6) is convergent if for any vectors

$$\hat{Y}_r^{(n_r-\nu)}(x,h), \qquad r = 1(1)q, \quad \nu = 1(1)n_r \ ,$$

chosen so that

$$\lim_{h\to 0} \left\| \hat{Y}_r^{(n_r-\nu)}(x,h) - Y_r^{(n_r-\nu)}(x,0) \right\| = 0 \ , \quad r = 1(1)q, \quad \nu = 1(1)n_r \ ,$$

---

* Stability and convergence are defined in the next section.

and any point $\bar{x} \in I$, then the solution of (2.6) with $h = \frac{\bar{x} - x}{m}$ at $\underline{x} + mh\underline{s}$ is such that

$$\lim_{h \to 0} \left\| \hat{Y}_r^{(n_r-\nu)}(x+mh, h) - Y_r^{(n_r-\nu)}(x+mh, 0) \right\| = 0 ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

A numerical solution which is "globally" close to the actual solution, must certainly be "locally" close to it, and the (local)* error at the k-th step is defined by

$$\psi_r^{[\nu]}(k) = Y_r^{(n_r-\nu)}(x+kh,h) - A_r^{[\nu,0]} Y_r^{(n_r-\nu)}(x+(k-1)h, h)$$

$$- \sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} Y_r^{(n_r-\nu+\sigma)}(x+kh, h)$$

$$- \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]} \bar{f}_r(\underline{x}+kh\underline{s}; Y(x+kh, h)) , \tag{2.7}$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

If $\bar{p}$ terms of Taylor series expansions for $\underline{Y}(x+kh, h)$ and $\hat{\underline{Y}}(x+kh, h)$ are to agree (for $\underline{Y}(x+(k-1)h,h) = \hat{\underline{Y}}(x+(k-1)h,h)$), it is necessary that $\psi_r^{[\nu]}(k)$ be of order $\bar{p}$. As $\underline{Y}(t,h)$ satisfies (1.2), (2.3) gives

---

* The final component of $\psi_r^{[\nu]}(k)$ is referred to as the truncation error.

$$\psi_r^{[\nu]}(k) = \sum_{\tau=0}^{\bar{p}} \frac{h^\tau}{\tau!} \Big[M^\tau - A_r^{[\nu,0]} (M - I)^\tau - \sum_{\sigma=1}^{\nu} \binom{\tau}{\sigma} M^\sigma A_r^{[\nu,\sigma]} M^{\tau-\sigma} \Big] Y_r^{(n_r-\nu+\tau)}(x+kh,0) + O(h^{\bar{p}+1}) ,$$

$r = 1(1)q, \quad \nu = 1(1)n_r,$

since

$$Y_r^{(n_r-\nu)}(x + (k-1)h, \underline{\mu}h) \equiv Y_r^{(n_r-\nu)}(x+kh, (\underline{\mu}-\underline{s})h) ,$$

and

$$\binom{\tau}{\sigma} = \begin{cases} \dfrac{\tau!}{(\tau-\sigma)!\,\sigma!} , & \tau \geq \sigma , \\ 0 , & \tau < \sigma . \end{cases}$$

For a convergent method, it is easily seen that $\psi_r^{[\nu]}(k)$ must at least be of order zero. Indeed, more is necessary and this leads to another definition.

<u>Definition (2.3)</u>: A method defined by (2,6) is consistent if there exists a diagonal matrix M' such that

$$(2.8) \quad \Big[M'^\tau - A_r^{[\nu,0]} (M' - I)^\tau - \sum_{\sigma=1}^{\nu} \binom{\tau}{\sigma} M^\sigma A_r^{[\nu,\sigma]} \Big] \underline{s} = 0 ,$$

$r = 1(1)q, \quad \nu = 1(1)n_r, \quad \tau = 0, 1.$

Now for a consistent method, the abscissae (and thus M) may be chosen so that the truncation error $\{\psi_r^{[\nu]}(k)\}$ is at least of order 1. For, suppose $M' \neq M$. Then conditions (2.8) for

$\tau = 0$ imply the conditions for $\tau = 1$ are valid with M' replaced by $M' + cI$ for any constant c. Then choosing c so that $(M' + cI)^{-1}$ exists, and thus rewriting $M A_r^{[\nu,1]}$ as $(M' + cI)\left[(M' + cI)^{-1} M A_r^{[\nu,1]}\right]$, it follows that (2.8) with $\tau = 1$ is valid for $M = M' = M' + cI$.

In practice, it is desirable to have a method of maximal order, and thus it is reasonable to choose the abscissae so that $M' = M$.

A second condition is necessary for convergence. Using (2.6) recursively, we obtain

$$(2.9)\quad \hat{Y}_r^{(n_r-\nu)}(x+mh,h) = \left[A_r^{[\nu,0]}\right]^m \hat{Y}_r^{(n_r-\nu)}(x,h)$$
$$+ \sum_{k=1}^{m} \left[A_r^{[\nu,0]}\right]^{m-k} \left[\sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} \hat{Y}_r^{(n_r-\nu+\sigma)}(x+kh,h)\right.$$
$$\left. + \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]} \bar{f}_r(\underline{x} + kh\underline{s};\ \underline{\hat{Y}}(x+kh,\ h))\right],$$
$$r = 1(1)q,\quad \nu = 1(1)n_r .$$

It is easily seen that local errors may grow if $\left\|\left[A_r^{[\nu,0]}\right]^m\right\|$ is not bounded, and this leads to another definition.

<u>Definition (2.4)</u>: A method defined by (2.6) is stable if there exists a constant α such that

$$\left\|\left[A_r^{[\nu,0]}\right]^m\right\| \leq \alpha , \qquad r = 1(1)q,\quad \nu = 1(1)n_r ,$$

for all m.

Butcher [12] shows that this condition is equivalent to the assumption that the eigenvalues of $A_r^{[\nu,o]}$ have magnitude not greater than unity, and that those of magnitude unity are simple. Strong stability is now defined for use in the existence theorem of the next section.

Definition (2.5): A method defined by (2.6) is strongly stable if it is stable, and only one eigenvalue has magnitude unity.

For first order systems, Butcher [12] shows with certain examples that a convergent method is necessarily stable and consistent. As his definitions of stability and consistency are equivalent to those given here, these examples may be used to show this result for the methods considered here. The generalization of Butcher's results is confined to showing that these conditions are sufficient for a method to be convergent.

Theorem (2.2): A stable and consistent method is convergent.

Proof: Here we consider a positive steplength h. (This does not restrict the proof in any way.)

As the choice of abscissae does not change the method, we may choose them so that $M = M'$. Then, using truncated Taylor expansions, the consistency conditions give

$$\psi_r^{[\nu]}(k) = \frac{h^2M^2}{2!} Y_r^{(n_r-\nu+2)}(x+kh, \underline{\theta}_1 h)$$

$$- A_r^{[\nu,0]} \frac{h^2(M-I)^2}{2!} Y_r^{(n_r-\nu+2)}(x+kh, (\underline{\theta}_2-\underline{s})h)$$

$$- hA_r^{[\nu,1]}\left[hMY^{(n_r-\nu+2)}(x+kh, \underline{\theta}_3 h)\right]$$

$$- \sum_{\sigma=2}^{\nu} \frac{h^\sigma}{\sigma!} A_r^{[\nu,\sigma]} Y_r^{(n_r-\nu+\sigma)}(x+kh, h)$$

$r = 1(1)q, \quad \nu = 1(1)n_r$,

where

$$|\theta_{ij}| \leq |\mu_i| , \quad i = 1(1)s+1, \quad j = 1,2,3.$$

As derivatives (2.1) are bounded in I, so also are the vectors above for $(x + kh + (\mu_i-1)h) \varepsilon I$, $i = 1(1)s+1$. Thus there exists a constant C such that

$$\left\|\psi_r^{[\nu]}(k)\right\| \leq Ch^2 , \tag{2.10}$$

$$r = 1(1)q, \quad \nu = 1(1)n_r ,$$

provided that $(x + kh + (\mu_i - 1)h) \varepsilon I$, $i = 1(1)s+1$.

For any $\bar{x} \varepsilon I$, $(\bar{x} > x)$, set $h = \frac{\bar{x}-x}{m}$ for some positive integer m, and choose $\hat{\underline{Y}}(x,h)$ so that

$$\lim_{h \to 0} \left\| \hat{Y}_r^{(n_r-\nu)}(x,h) - Y_r^{(n_r-\nu)}(x,h) \right\| = 0 ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

Define certain errors* by

$$E_r^{[\nu]}(k) = \hat{Y}_r^{(n_r-\nu)}(x+kh,h) - Y_r^{(n_r-\nu)}(x+kh,h)$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad k = 1(1)m .$$

Using (2.6) and (2.7) we obtain

$$(2.11) \quad E_r^{[\nu]}(k) = A_r^{[\nu,0]} E_r^{[\nu]}(k-1) + \sum_{\sigma=1}^{\nu-1} \frac{h^\sigma}{\sigma!} M^\sigma A_r^{[\nu,\sigma]} E_r^{[\nu-\sigma]}(k)$$

$$+ \frac{h^\nu}{\nu!} M^\nu A_r^{[\nu,\nu]} W_r(k) + \psi_r^{[\nu]}(k) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad k = 1(1)m ,$$

where

$$W_r(k) = \bar{f}_r(\underline{x} + kh\underline{s}; \hat{\underline{Y}}(x + kh,h)) - \bar{f}_r(\underline{x} + kh\underline{s}; \underline{Y}(x + kh,h)) .$$

To simplify the remainder of the argument, define N(s+1)-dimensional column vectors by their transposes.

$$\underline{E}(k)^T = (E_1^{[1]}(k)^T, \ldots, E_1^{[n_1]}(k)^T; \ldots; E_q^{[1]}(k)^T, \ldots, E_q^{[n_q]}(k)^T) ,$$

$$\underline{\psi}(k)^T = (\psi_1^{[1]}(k)^T, \ldots, \psi_1^{[n_1]}(k)^T; \ldots; \psi_q^{[1]}(k)^T, \ldots, \psi_q^{[n_q]}(k)^T) ,$$

and

$$\underline{W}(k)^T = (W_1(k)^T, \ldots, W_1(k)^T; \ldots; W_q(k)^T, \ldots, W_q(k)^T) ,$$

---

* The final components of $E_r^{[\nu]}(k)$ are referred to as accumulated errors.

and matrices A, which is block diagonal, $A'(h)$ and $A^*(h)$ which have powers of h in off-diagonal blocks, so that (2.11) may be written

$$\underline{E}(k) = A\,\underline{E}(k-1) + h\,A'(h)\underline{E}(k) + h\,A^*(h)\underline{W}(k) + \underline{\psi}(k) .$$

Using this recursively, we obtain

$$\underline{E}(m) = A^m\underline{E}(0) + h \sum_{k=1}^{m} A^{m-k}(A'(h)\underline{E}(k) + A^*(h)\underline{W}(k)) + \sum_{k=1}^{m} A^{m-k}\underline{\psi}(k) .$$

By definition, $\bar{f}_r(\underline{x};\underline{Y}(x,h))$ satisfies a Lipschitz condition on $\underline{Y}$ with constant L, and thus

$$\|\underline{W}(k)\| \leq L\,\|\underline{E}(k)\| .$$

Further taking norms this leads to

$$\|\underline{E}(m)\| \leq \|A^m\|\,\|\underline{E}(0)\| + h \sum_{k=1}^{m} \left[\|A^{m-k}\,A'(h)\| + L\,\|A^{m-k}\,A^*(h)\|\right]\|\underline{E}(k)\| + \sum_{k=1}^{m} \|A^{m-k}\|\,\|\underline{\psi}(k)\| .$$

Now for $\|A'(h)\| < \gamma$ and $\|A^*(h)\| < \delta$, and h sufficiently small, stability and (2.10) give

$$\|\underline{E}(m)\| \leq \alpha\,\|\underline{E}(0)\| + h \sum_{k=1}^{m} (\alpha\gamma + \alpha\delta L)\,\|\underline{E}(k)\| + m\alpha C h^2 .$$

Now define $\varepsilon_o = \alpha\,\|\underline{E}(0)\|$, and

$$\varepsilon_m = \varepsilon_o + h \sum_{k=1}^{m} (\alpha\gamma + \alpha\delta L)\varepsilon_k + m\alpha C h^2 , \qquad m = 1,2, \ldots .$$

Then for h sufficiently small, $\varepsilon_m \geq \|\underline{E}(m)\|$; for

$$\varepsilon_m - \|\underline{E}(m)\| \geq h \sum_{k=1}^{m} (\alpha\gamma + \alpha\delta L)(\varepsilon_k - \|\underline{E}(k)\|)$$

is possible for small h only if the left side is positive, and

the result follows by induction on m .

Now

$$\varepsilon_m - \varepsilon_{m-1} = h\,\alpha(\gamma + \delta_L)\varepsilon_m + \alpha\, C\, h^2 ,$$

so that

$$(\varepsilon_m + \tfrac{hC}{\gamma+\delta_L}) = (1 - h\,\alpha(\gamma + \delta_L))^{-1} (\varepsilon_{m-1} + \tfrac{hC}{\gamma+\delta_L})$$

$$= (1 - h\,\alpha(\gamma+\delta_L))^{-m} (\varepsilon_o + \tfrac{hC}{\gamma+\delta_L}) .$$

If h is chosen so that (2.6) defines a contraction mapping, and $h \cdot \alpha(\gamma + \delta_L) < 1$, then for all m,

$$(1 - h\,\alpha(\gamma + \delta_L))^{-m} \leq \exp\left(\frac{m\, h\, \alpha(\gamma + \delta_L)}{1 - h\,\alpha(\gamma + \delta_L)}\right) ,$$

and thus

$$\|E(m)\| \leq \varepsilon_m \leq \alpha \|E(0)\| \exp\left(\frac{(\bar{x}-x)\,\alpha(\gamma + \delta_L)}{1 - h\,\alpha(\gamma + \delta_L)}\right)$$

$$+ \frac{(\bar{x}-x)C}{m(\gamma + \delta_L)} \exp\left(\frac{(\bar{x}-x)\,\alpha(\gamma + \delta_L)}{1 - h\,\alpha(\gamma + \delta_L)}\right) .$$

By the choice of initial conditions, the right hand side converges to zero as m increases to infinity, and convergence is proved.

6. An Existence Theorem

Here it is proved that there exists a solution to a first order system of differential equations using a sequence of approximations generated by one of a certain general class of numerical methods. The result extends immediately to a system of equations of arbitrary orders by reduction to a first order system as in Chapter I. Initially a lemma is proved.

Lemma (2.1): If a numerical method is strongly stable, and consistent, and $\underline{y}_k$, $k = 1,2,3,\ldots$, is a sequence of uniformly bounded vectors of R_{s+1}, then there exists a constant K such that

$$\left\| \sum_{k=1}^{m} (A_r^{[1,0]})^k \, (I - A_r^{[1,0]}) \, \underline{y}_k \right\| < K, \qquad r = 1(1)q,$$

for all m.

Proof: Each value of r is considered individually, and T_r is chosen so that

$$J_r = T_r^{-1} A_r^{1,0} T_r$$

is the Jordan canonical form of $A_r^{[1,0]}$. Then T_r^{-1} exists, and $T_r^{-1} \underline{y}_k$, $k = 1,2,3,\ldots$, is a uniformly bounded sequence. Further by (2.8) with $\tau = 0$, $A_r^{[1,0]}$ has an eigenvalue equal to 1, with eigenvector $\underline{s}$; thus J_r has the same eigenvalue with eigenvector $T_r^{-1}\underline{s}$. By definition, the remaining eigenvalues are of magnitude less than unity. Let these be represented by

λ_i of multiplicity k_i, and corresponding eigenvectors by g_{ij}, $i = 1(1)\bar{s}$, $j = 1(1)k_i$, $k_1 + \ldots + k_{\bar{s}} = s$, so that for each value of i

$$J_r g_{i1} = \lambda_i g_{i1} ,$$

$$J_r g_{ij} = \lambda_i g_{ij} + g_{ij-1} , \quad j = 2(1)k_i .$$

Let

$$T_r^{-1} \underline{v}_k = \sum_{i=1}^{\bar{s}} \sum_{j=1}^{k_i} a_{ijk} g_{ij} + a_k \underline{s} .$$

Since the vectors $T_r^{-1} \underline{v}_k$ are uniformly bounded, there exists $a > 0$ such that

$$|a_{ijk}| < a, \quad i = 1(1)\bar{s} , \quad j = 1(1)k_i$$

for all k. Thus

$$\left\| \sum_{k=1}^{m} (A_r^{[1.0]})^k (I - A_r^{[1,0]}) \underline{v}_k \right\| = \left\| T_r \sum_{k=1}^{m} J_r^k (I - J_r) T_r^{-1} \underline{v}_k \right\|$$

$$= \left\| T_r \sum_{k=1}^{m} J_r^k (I - J_r) \sum_{i=1}^{\bar{s}} \sum_{j=1}^{k_i} a_{ijk} g_{ij} \right\|$$

$$\left\| T_r \right\| s a \max_{i,j,k} \left\{ \left\| \sum_{k=1}^{m} J_r^k (I - J_r) g_{ij} \right\| \right\} ,$$

since $(I - J_r)\underline{s} = 0$. Further

$$\left\| \sum_{k=1}^{m} J_r^k g_{ij} \right\| = \left\| \sum_{k=1}^{m} \left[\lambda_i^k g_{ij} + \binom{k}{1} \lambda_i^{k-1} g_{ij-1} + \ldots + \binom{k}{j-1} \lambda_i^{k-j+1} g_{i1} \right] \right\|$$

$$\leq \sum_{k=1}^{m} \left[|\lambda_i|^k + \binom{k}{1} |\lambda_i|^{k-1} + \ldots + \binom{k}{j-1} |\lambda_i|^{k-j+1} \right] \max_j \left\| g_{ij} \right\| ,$$

and the final expression is bounded for all m (as may easily be shown by the ratio test for infinite series). Thus the result is proved.

Corollary 1: Under the conditions of the lemma, there exists a constant $\overline{K}$ such that

$$\left\| \sum_{k=1}^{m} \left((A_r^{[1,0]})^k - (A_r^{[1,0]})^m \right) \underline{y}_k \right\| \leq \overline{K} \quad , \quad r = 1(1)q,$$

for all m.

Proof: As in the lemma it follows that

$$\left\| \sum_{k=1}^{m} \left((A_r^{[1,0]})^k - (A_r^{[1,0]})^m \right) \underline{y}_k \right\| \leq \left\| T_r \right\| \text{ sa } \max_{ij} \left\| \sum_{k=1}^{m} (J_r^k - J_r^m) g_{ij} \right\| ,$$

and the result follows as before, since

$$\left\| \sum_{k=1}^{m} J_r^m g_{ij} \right\| \leq m \left[|\lambda_i|^m + \binom{m}{1} |\lambda_i|^{m-1} + \ldots + \binom{m}{j-1} |\lambda_i|^{m-j+1} \right] \max_j \left\| g_{ij} \right\|$$

and is bounded for all m (and, in fact, converges to zero as m tends to infinity).

Corollary 2: For any vector $\underline{y}$, $(A_r^{[1,0]})^m \underline{y}$, $m = 1, 2, \ldots$ forms a Cauchy sequence.

This follows immediately from Corollary 1.

Hull and Luxemburg [25] show that the Peano existence theorem can be proved only for strongly stable methods. Indeed, for the solution of $y' = \sqrt{y}$, $y(0) = 0$, use the formula $\hat{y}(x+2h) - \hat{y}(x) = 2h \sqrt{\hat{y}(x)}$ (which is not strongly stable)

with starting conditions $y(0) = 0$, $y(h) = h$. Then $y(nh) = 0$ for n even, $y(nh) \to t^2/4$ for n odd. In this case, the difference of two successive approximate values is not small, as required in the proof. The proof of the theorem below uses approximations generated by (2.6) for a strongly stable, consistent method, and is much the same as that given by Hull and Luxemburg [25] for multi-step methods.

Theorem (2.3): Let $f_r(t; \underline{y})$, $r = 1(1)q$, be continuous functions in an open region D of R_{q+1}. Then for any point $(x; \underline{y}_o)$ of D, there exists an interval I with $x \in I$ and functions $\phi_r(t)$, $r = 1(1)q$, with

$$\frac{d\phi_r(t)}{dt} = f_r(t; \underline{\phi}(t)), \quad r = 1(1)q \ ,$$

for $t \in I$, and

$$\phi_r(x) = y_{ro} \ , \qquad r = 1(1)q \ .$$

Proof: Since the functions are continuous in D, there exists a closed region $\bar{D}$ of D containing $(x; \underline{y}_o)$ in which a maximum is attained,

$$K_1 = \max_{r,(t;\underline{y}) \in \bar{D}} \left| f_r(t;\underline{y}) \right| \ .$$

Assume that $\hat{\underline{y}}(x,h)$ is determined so that

$$\left| \hat{y}_r(x + (\mu_i - 1)h) - y_{ro} \right| \leq K_2 h, \quad i = 1(1)s+1, \ r = 1(1)q,$$

where the steplength h is positive (this restriction is not

necessary), and K_2 is a constant. Further, let a and b be chosen so that $\bar{D}$ contains the neighbourhood N_1 of $(x;y_o)$:

$$|t - x| \leq a, \quad |y_r - y_{ro}| \leq b \quad ,$$

and define

$$d = \max\left\{\frac{b}{a},\ K_2,\ \alpha(K_2 + \delta K_1)\right\} \quad ,$$

where α and δ are defined in Theorem (2.2). Then we prove the theorem for $0 \leq t-x \leq a' = \frac{b}{d}$.

(i) Define a neighbourhood $\bar{N}_1$ in $R_{(s+1)(q+1)}$ by

$$(\underline{x};\ \underline{Y}(x,h)) \in \bar{N}_1 \quad \text{if} \quad (x_i;\ \underline{y}(x_i)) \in N_1 \quad .$$

Then approximations $(\underline{x} + kh\underline{s},\ \hat{\underline{Y}}(x+kh,\ h))$ lie in $\bar{N}_1$ for $kh \leq a'$. For, let $\bar{x}$ be a point with $\bar{x} - x = mh \leq a'$ for a positive integer m. Then (2.9) and the first consistency condition give

$$(2.12)\quad \hat{Y}_r(x+mh,h) - Y_r(x,0)$$

$$= (A_r^{[1,0]})^m(\hat{Y}_r(x,h) - Y_r(x,0)) + h \sum_{k=1}^{m} (A_r^{[1,0]})^{m-k} \cdot$$

$$MA_r^{[1,1]}\bar{f}_r(x+kh\underline{s};\ \hat{\underline{Y}}(x+kh,h)) \quad ,$$

$$r = 1(1)q.$$

It follows from the starting conditions and the definition of a' that $(\underline{x};\ \hat{\underline{Y}}(x,h)) \in \bar{N}_1$. Assume that $(\underline{x} + kh\underline{s},\ \hat{\underline{Y}}(x + kh,h)) \in \bar{N}_1$, $k = 0,1, \ldots, m-1$, and consider the iteration

$$\hat{Y}_r^{(0)}(x+mh,h) = Y_r(x,0) + b\,\underline{s}\ ,$$

$$\hat{Y}_r^{(\ell+1)}(x+mh,h) = Y_r(x,0) + (A_r^{[1,0]})^m(\hat{Y}_r(x,h) - Y_r(x,0))$$

$$+ h\sum_{k=1}^{m-1}(A_r^{[1,0]})^{m-k}\, M\, A_r^{[1,1]}\, \overline{f}_r(\underline{x}+kh\underline{s};\ \hat{\underline{Y}}(x+kh,h)) + h\, MA_r^{[1,1]}\cdot$$

$$\overline{f}_r(\underline{x}+mh\underline{s};\hat{Y}^{(\ell)}(x+mh,h)),$$

$r = 1(1)q,\quad \ell = 1,2,\ldots$.

Now $(\underline{x} + mh\underline{s};\ \underline{Y}^{(0)}(x + mh,h)) \in \overline{N}_1$ by definition of N_1, and assume this is the case for some value of ℓ. Then for $\ell+1$

$$\left\|\hat{Y}_r^{(\ell+1)}(x+mh,h) - Y_r(x,0)\right\|$$

$$\leq \left\|(A_r^{[1,0]})^m\right\| \left\|\hat{Y}_r(x,h) - Y_r(x,0)\right\| + h\sum_{k=1}^{m-1}\left\|(A_r^{1,0})^{m-k}\right\|\left\|MA_r^{[1,1]}\right\|\cdot$$

$$\left\|\overline{f}_r(\underline{x}+kh\underline{s};\ \hat{\underline{Y}}(x+kh,h)\right\|$$

$$+ h\left\|MA_r^{[1,1]}\right\|\left\|\overline{f}_r(\underline{x}+mh\underline{s};\ \hat{\underline{Y}}^{(\ell)}(x+mh,h))\right\|$$

$$\leq \alpha K_2 h + h(m-1)\alpha\delta K_1 + h\delta K_1 \leq \alpha(K_2 + \delta K_1)\, m\, h \leq b\ ,$$

(since $\alpha \geq 1$), implying that $(\underline{x} + mh\underline{s};\ \hat{\underline{Y}}^{(\ell+1)}(x+mh,h)) \in \overline{N}_1$.

Thus by induction on ℓ, the iteration gives an infinite sequence, every element of which lies inside the closed neighbourhood $\overline{N}_1$. By the Bolzano-Weierstress theorem, this sequence contains a convergent subsequence whose limit satisfies (2.12). By a further induction on m the result follows.

(ii) For a strongly stable method, the difference between two successive approximations is O(h). Indeed (2.9) gives

$$\hat{Y}_r(x+mh,h) - \hat{Y}_r(x+(m-1)h,h)$$

$$= (A_r^{[1,0]})^{m-1} (\hat{Y}_r(x+h,h) - \hat{Y}_r(x,h)) + h M A_r^{[1,1]} \bar{f}_r(\underline{x}+mh\underline{s}; \underline{\hat{Y}}(x+mh,h))$$

$$+ h \sum_{k=2}^{m} (A_r^{[1,0]})^{m-1-k} (A_r^{[1,0]} - I) M A_r^{[1,1]} \bar{f}_r(\underline{x}+kh\underline{s}; \underline{\hat{Y}}(x+kh,h))$$

$$- h (A_r^{[1,0]})^{m-k} M A_r^{[1,1]} \bar{f}_r(\underline{x}+h\underline{s}; \underline{\hat{Y}}(x+h,h)), \qquad r = 1(1)q .$$

Since $(\underline{x}+kh\underline{s}; \underline{\hat{Y}}(x+kh,h)) \in \bar{N}_1$, $k = 0(1)m$, the functions $\{\bar{f}_r\}$ are uniformly bounded; thus Lemma (2.1) may be applied to the third term of the right side. Using the bounds on the starting conditions for the first term, it follows that there exists K_3 independent of m such that

$$\left\| \hat{Y}_r(x+mh,h) - \hat{Y}_r(x+m-1)h,h) \right\| \leq K_3 h, \qquad r = 1(1)q.$$

For each value of m, we define vector functions $\underline{\psi}_m(t)$ by interpolation from the approximations $\underline{\hat{Y}}(x+kh,h)$, $k = 0(1)m$. Thus for every m, and any t, t' with $0 \leq t-x$, $t'-x \leq a'$, it follows that

$$\| \underline{\psi}_m(t) - \underline{\psi}_m(t') \| \leq K_3 |t-t'| .$$

(iii) Consider any sequence of h's which converges to zero. Then the corresponding vector functions $\underline{\psi}_m(t)$ form an equicontinuous set by (ii) which is uniformly bounded by (i). By extending Arzela's theorem (Kolmogorov and Fomin [26, p. 54]) for finite-dimensional vectors, there exists a subsequence $\{h_j\}$ such

that the corresponding sequence $\{\underline{\psi}_{m_j}(t)\}$ converges to a limit $\underline{\psi}(t)$ which is continuous for $0 \leq t-x \leq a'$. Now define* functions $\{\phi_r(t)\}$ by

$$\left[\psi_r(t)\right]_{s+1} = \phi_r(t), \qquad r = 1(1)q,$$

and we must show the vector function $\underline{\phi}(t)$ satisfies the given differential system.

(iv) First (2.12) may be written

$$\hat{Y}_r(x+mh,h) - Y_r(x,0)$$

$$= (A_r^{[1,0]})^m (\hat{Y}_r(x,h) - Y_r(x,0)) + h \sum_{k=1}^{m} \left[(A_r^{[1,0]})^{m-k} - (A_r^{[1,0]})^m \right] MA_r^{[1,1]} \cdot$$

$$\overline{f}_r(\underline{x}+kh\underline{s};\ \hat{\underline{Y}}(x+kh,h))$$

$$+ (A_r^{[1,0]})^m MA_r^{[1,1]}\ h \sum_{k=1}^{m} \overline{f}_r(\underline{x}+kh\underline{s};\ \hat{\underline{Y}}(x+kh,h)),$$

$$r = 1(1)q \ .$$

For the subsequence $\{h_j\}$ which converges to zero, it follows that the first term on the right side converges to zero by the stability of the method, and the choice of $\hat{Y}_r(x,h)$. By (ii) and Corollary 1, it follows that the second term converges to zero. Thus for any h_j of the subsequence in (iii), we may write

---

* By previous definitions and convergence, it follows that $\left[\psi_r(t)\right]_i = \left[\psi_r(t)\right]_j$, for all $i \neq j$.

$$\lim_{m_j \to \infty} \psi_{m_j r}(\bar{x})$$

$$= Y_r(x,0) + \lim_{m_j \to \infty} \left\{ \left(A_r^{[1,0]}\right)^{m_j} M A_r^{[1,1]} \left[h_j \sum_{k=1}^{m_j} \bar{f}_r(\underline{x}+kh_j\underline{s};\ \underline{\psi}_{m_j}(x+kh_j)) \right] \right\}.$$

Now by Corollary 2, $\lim_{m_j \to \infty} \left\{ \left(A_r^{[1,0]}\right)^{m_j} M A_r^{[1,1]} \right\}$ exists. Further each component of the expression in square brackets is a Riemann sum. As the sequence $\{\underline{\psi}_{m_j}\}$ converges uniformly, so also does the sequence of Riemann sums, so that

$$\lim_{m_j \to \infty} \left[h_j \sum_{k=1}^{m_j} \bar{f}_r(\underline{x} + kh_j\underline{s};\ \underline{\psi}_{m_j}(x+kh_j)) \right]$$

$$= \int_x^{\bar{x}} \bar{f}_r(t;\ \underline{\psi}(t))dt = \underline{s} \int_x^{\bar{x}} f_r(t;\ \underline{\phi}(t))dt$$

since each component of $\bar{f}_r(t;\ \underline{\psi}(t))$ is the same. Since the limit of a product is the product of the limits, this gives

$$\psi_r(\bar{x}) = Y_r(x,0) + \lim_{m \to \infty} \left[\left(A_r^{[1,0]}\right)^m M A_r^{[1,1]} \right] \lim_{m_j \to \infty} \left[h_j \sum_{k=1}^{m_j} \bar{f}_r(\underline{x}+kh_j\underline{s};\ \underline{\psi}_{m_j}(x+kh_j)) \right]$$

$$= Y_r(x,0) + \int_x^{\bar{x}} f_r(t;\ \underline{\phi}(t))dt \lim_{m \to \infty} \left[\left(A_r^{[1,0]}\right)^m M A_r^{[1,1]} \underline{s} \right].$$

Now consistency implies

$$\lim_{m \to \infty} \left\{ \left(A_r^{[1,0]}\right)^m M A_r^{[1,1]} \underline{s} \right\} = \lim_{m \to \infty} \left\{ \left(A_r^{[1,0]}\right)^m \left[M - A_r^{[1,0]}(M-I) \right] \underline{s} \right\}$$

$$= \lim_{m \to \infty} \left\{ \left[\left(A_r^{[1,0]}\right)^m - \left(A_r^{[1,0]}\right)^{m+1} \right] M\underline{s} + \underline{s} \right\} = \underline{s}$$

since $(A_r^{[1,0]})^m \underline{Mg}$ is a Cauchy sequence. Then for each component of $\psi_r(\bar{x})$ it follows that

$$\phi_r(\bar{x}) = y_{ro} + \int_x^{\bar{x}} f_r(t; \underline{\phi}(t)\,dt, \qquad r = 1(1)q,$$

completing the proof of the theorem.

CHAPTER III

EXPLICIT RUNGE-KUTTA METHODS

1. Introduction

Here we examine a class of explicit single step methods. These were first examined systematically by Runge [35, 36] and Kutta [27]. Choose $\mu_1 = 0$. From given or previously calculated approximations to the solution and its derivatives for a single point x, corresponding values are determined by means of a finite sequence of function evaluations at the points $x_i = x + \mu_i h$, $i = 1(1)\ s+1$; the values for $i = s+1$ give the required approximations at $x + h$.

In the previous matrix notation, the elements of each matrix $A_r^{[\nu,0]}$ are zero everywhere except in the final column, matrices $A_r^{[\nu,\sigma]}$, $0 < \sigma < \nu$, are lower triangular, and matrices $A_r^{[\nu,\nu]}$ are strictly lower triangular. If the methods are derived to satisfy (3.5) and (3.6) below for $p \geq 1$, it may be shown that such methods are stable and consistent, and therefore convergent. Methods considered in this chapter are further restricted to a class for which only the first columns of $A_r^{[\nu,\sigma]}$, $0 < \sigma < \nu$, are non-zero. More general methods may be considered as a sub-class of hybrid methods examined in Chapter VI.

For some given accuracy, it is desirable to obtain the required approximations with a minimum of work. In general, explicit Runge-Kutta methods require more function evaluations per step than, for example, multi-step methods of the same order.

However, these methods have the virtue of being self-starting, and thus may be used, for example, to start multi-step methods. Further, they may be used easily for the complete solution of problems in which non-uniform behaviour may be dealt with more economically using a variable steplength.

Within any class of methods, numerical results indicate that methods of high order with a large stepsize are most economical. As function evaluations often consume a large portion of the computing time, the number of function evaluations may be considered as a measure of the work done. Indeed, s-stage explicit Runge-Kutta methods require s function evaluations at each step for each equation, and Butcher [10] considers the problem of minimizing s for methods of fixed order.

Henrici [23, p. 102] , in an example, states that there exist explicit Runge-Kutta methods of order p, p arbitrary, with $s = (p!)^2$ stages. A better result of this nature, due to Cooper [19], is given in Theorem (3.1). For a method of either type, s is unreasonably large, and we attempt to improve on this result.

Taylor series expansions of (2.6) lead to non-linear constraints on the parameters. There are fewer parameters than constraints, and as the order of a method increases so also does the complexity of the constraints. Butcher [7] indicates relations among the constraints, and thus develops efficient methods of order six. Here we write down conditions which are sufficient for a method to be of order p, and propose a class of methods of (arbitrary) even orders. Exploiting the relations among the

constraints leads to methods of order eight. Although it appears that methods of higher orders in this class do not exist, the techniques may be used to propose similar classes which may contain methods of arbitrary orders.

2. Parameter Constraints

The restrictions discussed in the introduction lead to a simpler expression of explicit Runge-Kutta methods. Assuming that

$$\hat{y}_r^{(n_r-\nu)}(x) = y_r^{(n_r-\nu)}(x) + O(h^{p+1}),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r,$$

(2.6) may be written

$$\hat{y}_r^{(n_r-\nu)}(x_i) = \sum_{\tau=0}^{\nu-1} \frac{(\mu_i h)^\tau}{\tau!}\, \hat{y}_r^{(n_r-\nu+\tau)}(x) + \frac{(\mu_i h)^\nu}{\nu!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu]}\, \hat{y}_r^{(n_r)}(x_j),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1,$$

$$\hat{y}_r^{(n_r)}(x_i) = f_r(x_i ; \{\hat{y}_\rho^{(n_\rho-m)}(x_i)\}),$$

$$r = 1(1)q, \quad i = 1(1)s,$$

with $\mu_1 = 0$.

We define certain errors by

$$\psi_{ri}^{[\nu]} = \sum_{\tau=0}^{\nu-1} \frac{(\mu_i h)^\tau}{\tau!}\, y_r^{(n_r-\nu+\tau)}(x) + \frac{(\mu_i h)^\nu}{\nu!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu]}\, y_r^{(n_r)}(x_j) - y_r^{(n_r-\nu)}(x_i),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1,$$

and a Taylor series expansion gives (with a change of index)

$$\psi_{ri}^{[\nu]} = \sum_{\tau=0}^{p-\nu} \frac{\mu_i^{\nu} h^{\tau+\nu}}{(\tau+\nu)!} \left[\binom{\tau+\nu}{\nu} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu]} \mu_j^{\tau} - \mu_i^{\tau} \right] y_r^{(n_r+\tau)}(x) + O(h^{p+1}) .$$

It is also convenient to define*

$$\varepsilon_{ri}^{[\nu]} = \hat{y}_r^{(n_r-\nu)}(x_i) - y_r^{(n_r-\nu)}(x_i), \qquad \nu = 1(1)n_r ,$$

and

$$\varepsilon_{ri} = f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i)\}) - f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i)\}),$$

$$r = 1(1)q, \quad i = 1(1)s .$$

Thus it follows that

$$(3.1) \quad \varepsilon_{ri}^{[\nu]} = \psi_{ri}^{[\nu]} + \frac{(\mu_i h)^{\nu}}{\nu!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu]} \varepsilon_{rj} + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1,$$

and further that

$$\varepsilon_{ri} = f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i) + \psi_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{i-1} \lambda_{\rho ij}^{[m]} \varepsilon_{\rho j} + O(h^{p+1})\})$$

$$- f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i)\}) ,$$

$$r = 1(1)q, \quad i = 1(1)s.$$

We assume the existence of first and second partial derivatives of the functions, and define

---

* We refer to $\psi_{r\,s+1}^{[\nu]}$ as the truncation errors, and $\varepsilon_{r\,s+1}^{[\nu]}$ as the accumulated errors.

$$f_r(x_i)_{\rho m} = \left[\frac{\partial f_r(t;\ \{z_\eta^{[\nu]}\})}{\partial z_\rho^{[m]}}\right]_{(x_i\ ;\ \{y_\eta^{(n_\eta-\nu)}(x_i)\})} .$$

Then a Taylor expansion truncated after second derivatives gives

$$(3.2)\quad \varepsilon_{ri} = \sum_{\rho=1}^{q}\sum_{m=1}^{n_\rho} f_r(x_i)_{\rho m}\left[\psi_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{i-1}\lambda_{\rho ij}^{[m]}\varepsilon_{\rho j}\right] + \frac{E_{ri}}{2} + O(h^{p+1}),$$

$$r = 1(1)q, \quad i = 1(1)s ,$$

where E_{ri} is a sum of products of second order partial derivatives with squares of errors $\{\varepsilon_{\rho i}^{[m]}\}$. Thus if

$$\varepsilon_{\rho i}^{[m]} = O(h^{\xi_i}), \quad \rho = 1(1)q, \quad m = 1(1)n_\rho ,$$

it follows that

$$E_{ri} = O(h^{2\xi_i}), \quad r = 1(1)q, \quad i = 1(1)s+1.$$

To show that $\varepsilon_{\rho i}^{[m]} = O(h^{\xi_i})$, it is necessary to show that $\psi_{\rho i}^{[m]} = O(h^{\xi_i})$, and that $\varepsilon_{rj} = O(h^{\xi_i - m})$ if $\lambda_{rij}^{[m]} \neq 0$.

By the definition in Chapter II, a method is of order p if

$$(3.3)\quad \varepsilon_{r\,s+1}^{[\nu]} = O(h^{p+1}), \quad r = 1(1)q, \quad \nu = 1(1)n_r .$$

It is now shown that there exist methods of arbitrarily high order, and it is convenient to begin with a lemma.

Lemma (3.1): For distinct non-zero abscissae $\mu_{k+1}, \ldots, \mu_\ell$, and for each i, $i > \ell > 0$, the parameters $\lambda_{rij}^{[\nu]}$, $j = 1, k+1(1)\ell$, may be chosen so that

$$\lambda_{ril}^{[\nu]}\mu_1 + \sum_{j=k+1}^{\ell} \lambda_{rij}^{[\nu]} \mu_j^{\tau} = \mu_i^{\tau} \binom{\tau+\nu}{\tau}^{-1} ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad \tau = 0(1)\ell-k.$$

Proof: For each set of values of i, r and ν, we have $\ell-k+1$ linear equations in $\ell-k+1$ unknowns. The matrix of the system has a van der Monde determinant (since $\mu_1 = 0$) which is non-singular, and the result follows.

Corollary: It follows immediately from the lemma that the parameters $\lambda_{rij}^{[\nu]}$, $j = 1, k+1(1)i-1$, may be chosen so that

$$\lambda_{ril}^{[\nu]} \mu_1^{\tau} + \sum_{j=k+1}^{i-1} \binom{\tau+\nu}{\tau} \lambda_{rij}^{[\nu]} \mu_j^{\tau} = \mu_i^{\tau} ,$$

$$\tau = 0(1)\ell-k , \quad \ell < i ,$$

and thus so that

$$\psi_{ri}^{[\nu]} = O(h^{\ell-k+2}) , \qquad \nu = 1(1)n_r,$$

since $\nu \geq 1$.

Theorem (3.1): For any positive integer p, there exist explicit Runge-Kutta methods of order p.

Proof: A method with parameters satisfying (3.3) is constructed.

Consider abscissae $\mu_1,\ldots,\mu_{s+1}$ with $\mu_1 = 0$, $\mu_{s+1} = 1$. These are chosen so that they may be partitioned sequentially into groups G_k, $k = 1(1)p+1$, with $G_1 = \{\mu_1\}$, G_k, $k = 2(1)p$, having k-1 distinct non-zero abscissae, and $G_{p+1} = \{\mu_{s+1}\}$. Thus for an index i, $1 \le i \le s+1$, there corresponds an abscissa μ_i which lies in exactly one group G_k for some index k.

Now set $\lambda_{rij}^{[\nu]} = 0$ if $\mu_i \in G_k$ and $\mu_j \notin G_1 \cup G_{k-1} \cup G_k$. Then the corollary implies that the remaining parameters may be chosen so that for each value of i,

$$(3.4)\quad \psi_{r,i}^{[\nu]} = O(h^k), \quad \mu_i \in G_k, \quad r = 1(1)q, \quad \nu = 1(1)n_r.$$

Now as $\mu_i = 0$, the initial assumptions imply

$$\varepsilon_{r1}^{[\nu]} = O(h^{p+1}), \qquad r = 1(1)q, \quad \nu = 1(1)n_r,$$

and thus

$$\varepsilon_{r1} = O(h^{p+1}), \qquad r = 1(1)q\ .$$

For $\mu_i \in G_2$, the result for ε_{r1} and (3.4) substituted in (3.1) give

$$\varepsilon_{ri}^{[\nu]} = O(h^2), \qquad r = 1(1)q, \quad \nu = 1(1)n_r.$$

In (3.2) these give

$$\varepsilon_{ri} = O(h^2), \qquad r = 1(1)q.$$

Assume for $k > 2$, $\mu_i \in G_{k-1}$, that

$$\varepsilon_{ri}^{[\nu]} = O(h^{k-1}), \qquad r = 1(1)q, \quad \nu = 1(1)n_r\ ,$$

and

$$\varepsilon_{ri} = O(h^{k-1}), \qquad r = 1(1)q.$$

Then we prove the result for k. Suppose $\underline{i}$ is the first index with $\mu_{\underline{i}} \in G_k$. Then $\lambda_{\underline{i}j} \neq 0$ only if $\mu_j \in G_1 \cup G_{k-1}$. Then the result for ε_{rj} and (3.4) substituted in (3.1) give

$$\varepsilon_{r\underline{i}}^{[\nu]} = O(h^k), \qquad r = 1(1)q, \quad \nu = 1(1)n_r.$$

In (3.2) these give

$$\varepsilon_{r\underline{i}} = O(h^k), \qquad r = 1(1)q.$$

Similar results follow by induction for all i with $\mu_i \in G_k$. By induction on k, it follows that

$$\varepsilon_{ri}^{[\nu]} = O(h^k), \quad \mu_i \in G_k, \qquad r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 2(1)s+1.$$

Thus as $\mu_{s+1} \in G_{p+1}$,

$$\varepsilon_{r\,s+1}^{[\nu]} = O(h^{p+1}), \qquad r = 1(1)q, \quad \nu = 1(1)n_r.,$$

which is the required result.

It follows easily that a method of this type requires $s = \frac{1}{2}(p^2 - p+4)$ stages; thus a method of order six requires at most 17 stages. As Butcher [7] derives 7-stage sixth order methods, these methods are very inefficient. Indeed some improvement is possible for $p \leq 8$*.

---

* Shanks [38] develops a 9-stage method of order seven, and a 12-stage method of order eight.

Define Taylor expansions for the first partial derivatives by

$$f_r(x_i)_{\rho m} = \sum_{\tau=0}^{p} \frac{(\mu_i h)^{\tau}}{\tau!} f_r^{(\tau)}(x)_{\rho m} + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad i = 1(1)s, \quad \rho = 1(1)q, \quad m = 1(1)n_\rho .$$

Lemma (3.2): It is sufficient that the following conditions be satisfied for a method to be of order p.

$$(3.5) \quad \sum_{i_o=1}^{s} a_{r_o i_o}^{[m_o]} \mu_{i_o}^{\tau_1+m_1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+m_k} \left[\binom{m_k+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} \lambda_{r_k i_{k-1} i_k}^{[m_k]} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right] = 0 ,$$

for all non-negative integral k with

$$\tau_1 + \ldots + \tau_{k+1} + m_o + \ldots + m_k \le p, \quad \tau_j \ge 0, \quad m_j > 0,$$

and

$$(3.6) \quad \sum_{i_o=1}^{s} a_{r_o i_o}^{[m_o]} \mu_{i_o}^{\tau_1+m_1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+m_k} \left[\lambda_{r_k i_{k-1} i_k}^{[m_k]} \right] = 0 ,$$

$$\tau_1 + \ldots + \tau_k + m_o + \ldots + m_k + 2\xi_{i_k} \le p ,$$

where

$$\varepsilon_{r_k i_k}^{[m_k]} = O(h^{\xi_{i_k}}) , \qquad m_k = 1(1)n_{r_k} .$$

Proof: Using (3.2) recursively, and expanding first partial derivatives in Taylor series, the errors ε_{ri} may be expressed in

powers of h. For a method to be of order p, it is sufficient that in the corresponding power expansions of the accumulated errors given by (3.1), the coefficients of powers of h with index less than $p+1$ be zero. In each expansion the coefficient of each power of h is a sum of terms over the number $(k+1)$ of parameters $\lambda_{rij}^{[\nu]}$ occuring in each product, and the number of combinations of τ_i, $i = 1(1)k+1$, whose sum is equal to the index of the power. It is sufficient to show that each term is zero; two types of term arise from the recursion:

$$\frac{(h)^{m_o}}{m_o!} \sum_{i_o=1}^{s} a_{r_o i_o}^{[m_o]} \frac{(\mu_{i_o} h)^{\tau_1}}{\tau_1!} f_{r_o}^{(\tau_1)}(x)_{r_1 m_1} \frac{(\mu_{i_o} h)^{m_1}}{m_1!} \sum_{i_1=1}^{i_o-1} \lambda_{r_1 i_o i_1}^{[m_1]} \cdots$$

$$\cdots \sum_{i_{k-1}=1}^{i_{k-2}-1} \lambda_{r_{k-1} i_{k-2} i_{k-1}}^{m_{k-1}} \frac{(\mu_{i_{k-1}} h)^{\tau_k}}{k!} f_{r_{k-1}}^{(\tau_k)}(x)_{r_k m_k} \left[\psi_{r_k i_{k-1}}^{[m_k]}\right],$$

and

$$\frac{(h)^{m_o}}{m_o!} \sum_{i_o=1}^{s} a_{r_o i_o}^{[m_o]} \frac{(\mu_{i_o} h)^{\tau_1}}{\tau_1!} f_{r_o}^{(\tau_1)}(x)_{r_1 m_1} \frac{(\mu_{i_o} h)^{m_1}}{m_1!} \sum_{i_1=1}^{i_o-1} \lambda_{r_1 i_o i_1}^{[m_1]} \cdots$$

$$\cdots \sum_{i_{k-1}=1}^{i_{k-2}-1} \lambda_{r_{k-1} i_{k-2} i_{k-1}}^{[m_{k-1}]} \frac{(\mu_{i_{k-1}} h)^{\tau_k}}{\tau_k!} f_{r_{k-1}}^{(\tau_k)}(x)_{r_k m_k} \frac{(\mu_{i_{k-1}} h)^{m_k}}{m_k!} \cdot$$

$$\sum_{i_k=1}^{i_{k-1}-1} \lambda_{r_k i_{k-1} i_k}^{[m_k]} \frac{E_{r_k i_k}}{2} .$$

Now with the expression for $\psi_{ri}^{[\nu]}$, (3.5) implies that the first set of coefficients are zero for appropriate values of the indices. As $E_{r_k i_k} = O(h^{2\xi i_k})$, (3.6) implies that the second set of coefficients are zero for appropriate indices, and the result follows.

Suppose the recursive expansion giving the power series in h, has all coefficients independent of h. Then for a method to be of order p, it is necessary that each coefficient of h^{τ}, $\tau \leq p$, be zero. Further, these coefficients must be zero for every set of functions having appropriate Taylor series expansions. Again each coefficient is made up of a sum of terms (similar to those above). The functions for a general system of equations may be chosen so that all but one term are identically zero. Hence, it is necessary that each term in a coefficient be zero if a method is to be of order p.

Of the two types of term in the expansion above, only the first is independent of h. Thus by expanding E_{ri} in Taylor series about x, necessary conditions for a method to be of order p would be obtained. Indeed, the condition $\alpha_2 = 0$ as implied by (3.6) for methods of order greater than four, replaces the necessary condition given by Butcher [6].

$$(3.6') \quad [\,[\phi]^2] = \sum_{i=1}^{s} \sum_{j=1}^{i-1} a_i \mu_i^2 (\lambda_{ij}\mu_j)^2 = \frac{1}{20} .$$

Fifth order methods for a single differential equation with $\alpha_2 \neq 0$ are given by Cassity [14]. Such methods for a system are not known, and further, the restriction $\alpha_2 = 0$ for a fifth order method (and corresponding restrictions for higher order methods) leads to a simplification of parameter constraints given by (3.6). Indeed, a significant simplification of this type is reflected in the proof of Theorem (3.3) below.

3. Reduction to Methods for First Order Systems

The following result due to Cooper permits a simplification of the parameter constraints.

Theorem (3.2): Suppose the parameters are calculated so that

$$\lambda_{rij}^{[m]} = \lambda_{ij}^{[m]}, \qquad r = 1(1)q,$$

and

$$\mu_i \lambda_{ij}^{[m+1]} = \frac{m+1}{m}(\mu_i - \mu_j)\lambda_{ij}^{[m]} \qquad (3.7)$$

$$n = \max_r n_r , \quad m = 1(1)n-1, \quad i = 2(1)s+1, \quad j = 1(1)i-1.$$

Then writing λ_{ij} for $\lambda_{ij}^{[1]}$, (3.5) and (3.6) are satisfied if

$$\sum_{i_o=1}^{s} \alpha_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+1} \left[(\tau_{k+1}+1) \sum_{i_k=1}^{i_{k-1}-1} \lambda_{i_{k-1}i_k} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right] = 0 , \qquad (3.8)$$

$$\tau_1 + \ldots + \tau_{k+1} + (k+1) \leq p ,$$

and

$$\sum_{i_o=1}^{s} \alpha_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+1} \left[\lambda_{i_{k-1}i_k} \right] = 0 , \qquad (3.9)$$

$$\tau_1 + \ldots + \tau_{k+1} + (k+1) + 2\xi_{i_k} \leq p ,$$

where

$$(\tau+1)\quad \sum_{j=1}^{i_k-1} \lambda_{i_k j}\, \mu_j^{\tau} - \mu_{i_k}^{\tau} = 0, \qquad \tau = 0(1)\, \xi_{i_k} - 2.$$

Proof. (i) First for $m_j = 1$, $j = 0(1)k-1$, it is shown using (3.7) that (3.8) implies (3.5) for $m_k = 1,2,\ldots$. Indeed, for $m_k = 1$, these conditions are identical. Assume the result is valid for all values of $m_k < \gamma$, with $\gamma > 1$. Then for $m_k = \gamma$, we have

$$\sum_{i_o=1}^{s} a_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+\gamma} \left[\binom{\gamma+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} \lambda_{i_{k-1}i_k}^{[\gamma]} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right]$$

$$= \sum_{i_o=1}^{s} a_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+\gamma-1} \left[\binom{\gamma+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} \frac{\gamma}{\gamma-1} (\mu_{i_{k-1}} - \mu_{i_k}) \cdot \right.$$

$$\left. \lambda_{i_{k-1}i_k}^{[\gamma-1]} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right]$$

$$= \frac{\gamma+\tau_{k+1}}{\gamma-1} \left\{ \sum_{i_o=1}^{s} a_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+\gamma} \left[\binom{\gamma-1+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} \lambda_{i_{k-1}i_k}^{[\gamma-1]} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right] \right\}$$

$$- \frac{\tau_{k+1}+1}{\gamma-1} \left\{ \sum_{i_o=1}^{s} a_{i_o} \mu_{i_o}^{\tau_1+1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k+\gamma-1} \left[\binom{\gamma-1+\tau_{k+1}+1}{\tau_{k+1}+1} \cdot \right. \right.$$

$$\left. \left. \sum_{i_k=1}^{i_{k-1}-1} \lambda_{i_{k-1}i_k}^{[\gamma-1]} \mu_{i_k}^{\tau_{k+1}+1} - \mu_{i_{k-1}}^{\tau_{k+1}+1} \right] \right\}$$

$$= 0,$$

$$\tau_1 + \ldots + \tau_{k+1} + k + \gamma \le p .$$

Here, we have used the result with $\nu-1$ which implies (3.5) with

$$\tau_1 + \ldots + \tau_{k+1} + k + (\nu-1) \leq p.$$

Then the result follows by induction on m_k.

(ii) We must now show that the result is valid for any set of values of m_j, $j = 0(1)k-1$. In this case, for (3.5) we obtain

$$\sum_{i_o=1}^{s} {}^{[m_o]}\alpha_{i_o} \mu_{i_o}^{\tau_1^*+m_1} \sum_{i_1=1}^{i_o-1} \cdots \mu_{i_{k-1}}^{\tau_k^*+m_k} \left[\binom{m_k+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} {}^{[m_k]}\lambda_{i_{k-1}i_k} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right]$$

$$= \sum_{i_o=1}^{s} m_o \alpha_{i_o} (1-\mu_{i_o})^{m_o-1} \mu_{i_o}^{\tau_1^*+1} \sum_{i_1=1}^{i_o-1} \cdots m_{k-1} \lambda_{i_k-2 i_{k-1}} (\mu_{i_{k-2}} - \mu_{i_{k-1}})^{m_{k-1}-1}$$

$$\mu_{i_{K-1}}^{\tau_K^*+m_k} \left[\binom{m_k+\tau_{k+1}}{\tau_{k+1}} \sum_{i_k=1}^{i_{k-1}-1} {}^{[m_k]}\lambda_{i_{k-1}i_k} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}} \right] ,$$

$$\tau_1^* + \ldots + \tau_k^* + \tau_{k+1} + m_o + \ldots + m_k \leq p .$$

Since (3.5) is valid for $m_j = 1$, $j = 0(1)k-1$, and all values of m_k such that $\tau_j = \tau_j^*(1)\,\tau_j^* + m_{j-1} - 1$, $j = 1(1)k$, with

$$\tau_1 + \ldots + \tau_k + \tau_{k+1} + k + m_k \leq p ,$$

it follows easily that the right side is zero, and thus (3.5) is valid for all values of m_j, $j = 0(1)k$.

(iii) Using (3.7) and a proof similar to (ii), it follows that (3.9) implies (3.6) for the stated values of the indices.

Thus from a method for a first order system of equations, the recursion (3.7) may be used to generate a method of the same order for a system of differential equations of arbitrary orders. In Chapter VI, generalizations of (3.7) lead to a similar result for more general classes of methods.

4. Solution of the Parameter Constraints.

Here it is necessary only to consider methods for a first order system, and a method is conveniently expressed in the form:

$$\begin{array}{c|cccccc} \mu_1 & & & & & & \\ \mu_2 & \lambda_{21} & & & & & \\ \vdots & \vdots & \ddots & & & & \\ \mu_s & \lambda_{s1} & \cdots & \cdots & \cdots & \lambda_{s\,s-1} & \\ \hline \mu_{s+1} & \alpha_1 & \cdots & \cdots & \cdots & \alpha_{s-1} & \alpha_s \end{array}$$

Butcher [7] associates the following sixth order explicit Runge-Kutta method with an implicit method of the same order [9]. Certain features of this method help to clarify this association. Indeed the abscissae are chosen (for either method) from the quadrature points of a 4-point Lobatto quadrature, and the non-zero parameters $\{\alpha_i\}$ are the quadrature weights. Further, the

0							
$\frac{5+\sqrt{5}}{10}$	1						
$\frac{5-\sqrt{5}}{10}$	$\frac{1+\sqrt{5}}{4}$	$\frac{3-\sqrt{5}}{4}$					
$\frac{5+\sqrt{5}}{10}$	$-\frac{2+\sqrt{5}}{2}$	$-\frac{\sqrt{5}}{2}$	$2+\sqrt{5}$				
$\frac{5-\sqrt{5}}{10}$	$\frac{3+\sqrt{5}}{12}$	0	$\frac{5+\sqrt{5}}{12}$	$\frac{2-\sqrt{5}}{6}$			
$\frac{5+\sqrt{5}}{10}$	$\frac{3-\sqrt{5}}{12}$	0	$\frac{5+4\sqrt{5}}{6}$	$\frac{5-\sqrt{5}}{12}$	$-\frac{1+\sqrt{5}}{2}$		
1	$\frac{1}{6}$	0	$-\frac{55+25\sqrt{5}}{12}$	$-\frac{25-7\sqrt{5}}{12}$	$5+2\sqrt{5}$	$\frac{5-\sqrt{5}}{2}$	
1	$\frac{1}{12}$	0	0	0	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

TABLE (3.1) A Method of order 6.

structure of this method is identical to that used in the proof of Theorem (3.1). Indeed, the abscissae may be grouped (as suggested by dotted lines) so that

$$(3.10)\quad \lambda_{ij} = 0, \quad \mu_i \in G_m, \quad \mu_j \notin G_1 \cup G_{m-1} \cup G_m \quad .$$

This leads to the proposal of a class of methods of arbitrary (even) orders. Consider abscissae $\mu_1, \ldots, \mu_{s+1}$, with $\mu_1 = 0$, $\mu_{s+1} = 1$. For a method of order p, $s+1 = \frac{1}{8}(p^2 + 2p + 16)$ abscissae are required. These are sequentially partitioned into $\bar{s}+1 = \frac{1}{2}(p+4)$ groups so that $G_1 = \{\mu_1\}$, G_m, $m = 2(1)\bar{s}$, has $k-1$ distinct non-zero abscissae, and $G_{\bar{s}+1} = \{\mu_{s+1}\}$. The abscissae are chosen from the quadrature points of an $\bar{s}$-point Lobatto quadrature, and the parameters $\{\alpha_i\}$ corresponding to abscissae in

G_1 and G_s are the appropriate weights; otherwise, $\alpha_i = 0$. The remaining parameters are chosen to satisfy (3.10) and

$$(3.11) \qquad \sum_{j=1}^{i-1} \lambda_{ij} \mu_j^{\tau} = \frac{\mu_i^{\tau}}{\tau+1}, \quad \mu_i \in G_m, \quad \tau = 0(1)m-2.$$

These relations do not completely define the parameters, and in attempting to satisfy the remaining conditions of (3.8) for p=4, the following fourth order method is easily obtained.

0				
$\frac{1}{2}$	1			
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$		
1	0	-1	2	
1	$\frac{1}{6}$	0	$\frac{4}{6}$	$\frac{1}{6}$

TABLE 3.2: A Method of Order 4

This method corresponds to a certain fourth order implicit method given by Butcher [9] in the same way that that of Table (3.1) does to the corresponding sixth order implicit method. Thus it seems reasonable to expect the existence of higher order methods of this class. We now proceed to show how to derive the parameters for methods of order eight.

Indeed choose s = 11. Based on a 5-point Lobatto quadrature (see Butcher [9]) select abscissae $\mu_1 = 0$, $\mu_{11} = 1$, $\mu_2, \ldots, \mu_{10}$ from $\frac{1}{2}$, $\frac{1}{2} \pm \frac{\sqrt{21}}{14}$ in groups G_m so that $G_1 = \{\mu_1\}$ and G_m, m = 2(1)5, has m-1 distinct abscissae. Further,

$a_1 = a_{11} = \frac{1}{20}$, $a_2 = \ldots = a_7 = 0$; a_8, a_9, a_{10} are chosen from $\frac{16}{45}$, $\frac{49}{180}$ corresponding to the choice of μ_8, μ_9, μ_{10} respectively.

It is convenient to adopt the notation

$$\underline{\lambda}_{ij} = \sum_{\mu_\ell = \mu_j} \lambda_{i\ell}$$

for all values of the indices.

Further define (m-1)-dimensional column vectors, $v_{m\ell}$, where for $\mu_{i_o} \in G_m$, the $\tilde{i}_o$-th component* is given by

$$\left[v_{m\ell}\right]_{\tilde{i}_o} = \sum_{i_1=1}^{i_o-1} \lambda_{i_o i_1} \mu_{i_1} \cdots \left[\sum_{i_\ell=1}^{i_{\ell-1}-1} \lambda_{i_{\ell-1} i_\ell} \mu_{i_\ell}^{m-\ell} - \frac{\mu_{i_{\ell-1}}^{m-\ell}}{(m-\ell+1)} \right]$$

$$m = 3, 4, \qquad \ell = 1(1)m-1 \ .$$

Also certain $(m-1)x(m-1)$ matrices A_m are required in the proof of Theorem (3.3). For $\mu_j \in G_m$, the $\tilde{j}$-th column of A_m is defined by

$$\left[A_m\right]_{\tilde{j}} = \begin{vmatrix} \sum_{i_o=8}^{11} a_{i_o} \mu_{i_o}^2 & \sum_{i_1=5}^{7} \lambda_{i_o i_1} \mu_{i_1} \lambda_{i_1 j} \\ \sum_{i_o=8}^{11} a_{i_o} \mu_{i_o}^2 & \sum_{i_1=5}^{7} \lambda_{i_o i_1} \mu_{i_1} \lambda_{i_1 j} \mu_j \end{vmatrix}, \qquad m = 3,$$

and

$$\left[A_m\right]_{\tilde{j}} = \begin{vmatrix} \sum_{i=8}^{11} a_i \mu_i^2 \lambda_{ij} \mu_j \\ \sum_{i=8}^{11} a_i \mu_i^2 \lambda_{ij} \mu_j^2 \\ \sum_{i=8}^{11} a_i \mu_i^3 \lambda_{ij} \mu_j \end{vmatrix}, \qquad m = 4$$

---

* For the first abscissa $\mu_{\underline{i}_o} \in G_m$, $\underline{i}_o = 1$. Indeed, define $i_o = i_o - \frac{1}{2}(m^2 - 3m + 4)$.

For $\mu_i \in G_4$, the $\tilde{i}$-th column of a 3 x 3 matrix B is defined by

$$\left[B \right]_{\tilde{i}} = \begin{bmatrix} \sum_{j=3}^{4} \lambda_{ij} \mu_j \lambda_{j2} \\ \mu_i \sum_{j=3}^{4} \lambda_{ij} \mu_j \lambda_{j2} \\ \sum_{j=3}^{4} \lambda_{ij} \mu_j^2 \lambda_{j2} \end{bmatrix}$$

With the abscissae and parameters $\{a_{ri}\}$ chosen above, consider the constraints placed on the remaining parameters.

(1) $\lambda_{ij} = 0, \quad j < i, \quad \mu_i \in G_5, \quad \mu_j \in G_2 \cup G_3,$

$\mu_i \in G_4, \quad \mu_j \in G_2 \ .$

(2) $(\tau+1) \quad \sum_{j=1}^{i-1} \lambda_{ij} \mu_j^{\tau} = \mu_i^{\tau} \qquad \mu_i \in G_m, \quad m = 2(1)5, \quad \tau = 0(1)m-2.$

(3) $\sum_{i=j+1}^{s} a_i \mu_i \lambda_{ij} = a_j(1-\mu_j), \qquad j = 8, 9, 10.$

(4) $\sum_{i=1}^{s} a_i \mu_i^{\tau_1+1} \sum_{j=1}^{i-1} \lambda_{ij} \mu_j^{\tau_2+1} \underline{\lambda}_{j\ell}$

$$= a_\ell \left\{ \frac{1}{(\tau_2+1)(\tau_1+\tau_2+2)} - \frac{\mu_\ell^{\tau_2+1}}{(\tau_1+1)(\tau_2+1)} + \frac{\mu_\ell^{\tau_1+\tau_2+2}}{(\tau_1+1)(\tau_1+\tau_2+2)} \right\},$$

$\ell = 8, \quad \tau_1 + \tau_2 \le 2, \quad \tau_1 > 0 \ .$

(5) There exist constants* a_{31}, a_{41} such that

$$v_{m1} = a_{m1}\, v_{mm-1}, \qquad m = 3,4 .$$

(6) $\det(B) = 0.$

Theorem (3.3): If the abscissae and weights are based on a 5-point Lobatto quadrature, and the parameters satisfy (1) - (6), then for $p = 8$, all but six of the conditions (3.8) and (3.9) are valid. Further, the unsatisfied conditions are interdependent (and, as seen later, are satisfied for certain orderings of the abscissae).

Proof: (i) Constraints (5) and (6) are not sufficiently general to permit a concise proof. Indeed, in part, each value of k for (3.8) must be considered individually.

$k = 0$: By Lobatto quadrature,

$$\text{(a)}\quad (\tau+1) \sum_{i=1}^{11} \alpha_i \mu_i = (\tau+1) \int_0^1 \mu\, d\mu = 1, \qquad \tau = 0(1)7 ,$$

whence (3.8) for $k = 0$.

Before proceeding, we show the validity of (3) for $j = 1(1)11$. Indeed it is valid for $j = 2,3,4,11$ trivially, and for $j = 8,9,10$ by assumption. By (2), for $\mu_i \in G_5$,

$$\text{(b)}\quad \sum_{i=8}^{11} \alpha_i \mu_i \left[\sum_{j=5}^{i-1} \lambda_{ij} \mu_j^{\tau} - \frac{\mu_i^{\tau}}{\tau+1} \right] = 0, \qquad \tau = 1,2,3.$$

and using (a), it follows that

---

* By definition of $v_{m\ell}$ it follows immediately that $v_{42} = \frac{a_{41}}{a_{31}}\, v_{43}$.

$$(c)\quad \sum_{j=8}^{10}\left[\sum_{i=8}^{11} a_i \mu_i \underline{\lambda}_{ij} - a_j(1-\mu_j)\right]\mu_j^{\tau} = 0,\ \tau = 1,2,3,$$

since μ_j, $j = 2(1)10$, can take only one of three distinct values. This system has a non-singular matrix of coefficients, and thus each term in square brackets is identically zero, so that

$$\sum_{i=8}^{11} a_i \mu_i \underline{\lambda}_{ij} = a_j(1-\mu_j), \qquad j = 8, 9, 10.$$

As μ_5, μ_6, μ_7 are distinct, (3) for $j = 8, 9, 10$ implies (3) for $j = 5, 6, 7$. Finally as

$$\sum_{j=1}^{10}\left[\sum_{i=j+1}^{11} a_i \mu_i \lambda_{ij}\right] = \tfrac{1}{2} = \sum_{j=1}^{10} a_j(1-\mu_j),$$

(3) is valid for $j = 1$.

Now (3.8) for $k > 0$ with $\tau_1 = 0$, becomes

$$\sum_{i_1=1}^{s} a_{i_1}(1-\mu_{i_1})\mu_{i_1}^{\tau_2+1} \cdots \mu_{i_{k-1}}^{\tau_k+1}\left[(\tau_{k+1}+1)\sum_{i_k=1}^{i_{k-1}-1}\lambda_{i_{k-1}i_k}\mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k-1}}^{\tau_{k+1}}\right] = 0$$

and this is valid if (3.8) is valid for k-1 for all values of τ_1. Thus we need only consider (3.8) for values of $k > 0$ with $\tau_1 > 0$.

k=1: As in (b) and (c), (2) for $\mu_i \in G_5$ may be used to show that

$$(d)\quad \sum_{j=8}^{10}\left[\sum_{i=8}^{11} a_i \mu_i^{\tau_1+1}\underline{\lambda}_{ij} - \frac{a_j}{\tau_1+1}(1-\mu_j^{\tau_1+1})\right]\mu_j^{\tau_2} = 0,$$

$$\tau_2 = 1, 2, 3, \qquad \tau_1 + \tau_2 + 1 \le 7.$$

Again the expression in square brackets is zero, and the value of τ_2, $\tau_2 > 0$, is immaterial. Using (a), (d) may be written in the form of (b) for $\tau_1 + \tau_2 \leq 6$, whence (3.8).

k=2: For $\mu_j \in G_4 \cup G_5$, (2) implies (3.8) for $\tau_3 = 0,1,2$. Similar to the derivation of (c) from (b) this gives

$$\sum_{\ell=8}^{10} \left[\sum_{i=8}^{11} a_i \mu_i^{\tau_1+1} \sum_{j=5}^{i-1} \lambda_{ij} \mu_j^{\tau_2+1} \lambda_{j\ell} - a_\ell \left\{ \frac{1}{(\tau_2+1)(\tau_1+\tau_2+2)} - \frac{\mu_\ell^{\tau_2+1}}{(\tau_1+1)(\tau_2+1)} + \frac{\mu_\ell^{\tau_1+\tau_2+2}}{(\tau_1+1)(\tau_1+\tau_2+2)} \right\} \right] \mu_\ell^{\tau_3} = 0,$$

$$\tau_3 = 1, 2, \quad \tau_1 + \tau_2 + 2 + \tau_3 \leq 7 .$$

For $\ell=8$, (4) asserts that the expression in square brackets is zero for $\tau_1 + \tau_2 \leq 2$, $\tau_1 > 0$. There remains a system of two equations in two unknowns, and as before the expression in square brackets is zero for $\ell = 9, 10$. Again the value of τ_3 is immaterial for these values of τ_1 and τ_2, whence (3.8) for all choices of the indices.

Further, it follows that

(e) $$A_4 \, v_{41} = 0 ,$$

and as $v_{41} \neq 0$, A_4 is singular (and further is of rank 2).

k=3: For $\mu_\ell \in G_3 \cup G_4 \cup G_5$, (2) implies (3.8) for $\tau_4 = 0,1$. Also, (e) and (5) imply $A_4 \, v_{42} = 0$ which gives (3.8) for

$\tau_1 + \tau_2 \leq 2$, $\tau_1 > 0$, $\tau_3 = 0$, $\tau_4 = 2$. Similarly after proving (f) (below) for $k = 4$, (5) will imply $A_3 v_{31} = 0$ which gives (3.8) for $\tau_1 = 1$, $\tau_2 = 0$, $\tau_3 = 1$, $\tau_4 = 2$. There remains only (3.8) for $\tau_1 = 1$, $\tau_2 = \tau_3 = 0$, $\tau_4 = 3$ which is not satisfied for all choices of the abscissae.

$k = 4$: For $\mu_m \in G_2 \cup G_3 \cup G_4 \cup G_5$, (2) implies (3.8) is valid for $\tau_5 = 0$. Again (e) and (5) imply $A_4 v_{43} = 0$ which gives (3.8) for $\tau_1 + \tau_2 \leq 2$, $\tau_1 > 0$, $\tau_3 = 0$, $\tau_4 = 0, 1$, $\tau_5 = 1$. This leads to

$$\sum_{i=8}^{11} a_i \mu_i^2 \sum_{j=5}^{i-1} \lambda_{ij} \mu_j^{\tau_2+1} \sum_{\ell=3}^{j-1} \lambda_{j\ell} \mu_\ell \lambda_{\ell 2} = 0, \quad \tau_2 = 0,1;$$

by (1) non-zero contributions occur only for $j = 5,6,7$, and thus (6) implies

$$\sum_{i=8}^{11} a_i \mu_i^2 \sum_{j=5}^{i-1} \lambda_{ij} \mu_j \sum_{\ell=3}^{j-1} \lambda_{j\ell} \mu_\ell^2 \lambda_{\ell 2} = 0 ,$$

giving (3.8) for $\tau_1 = 1$, $\tau_2 = 0$, $\tau_3 = 1$, $\tau_4 = 0$, $\tau_5 = 1$. Further

$$\text{(f)} \qquad A_3 v_{32} = 0,$$

and the proof for $k = 3$ may be completed. There remains only (3.8) for

$\tau_1 = 1$, $\tau_2 = \tau_3 = \tau_4 = 0$, $\tau_5 = 2$ which is not satisfied for all choices of the abscissae.

k = 5: Again (2) implies (3.8) for $\tau_6 = 0$. The condition with $\tau_1 = 1$, $\tau_2 = \ldots = \tau_5 = 0$, $\tau_6 = 1$, is not satisfied for all choices of the abscissae.

For k = 6, (2) implies (3.8). By the discussion for $\tau_1 = 0$, (3.8) is valid for all remaining values of the indices except those which would be otherwise derived from the three unsatisfied conditions.

(ii) To show (3.9) is valid, for each value of i_k we consider different values of k.

$i_k = 2$: $\xi_{i_k} = 2$

k = 0: $a_2 = 0$ implies (3.9).

k = 1: $a_2 = a_3 = a_4 = 0$, $\lambda_{i2} = 0$ for $i > 4$, implies (3.9).

k = 2: $a_2 = \ldots = a_7 = 0$, $\lambda_{j2} = 0$ for $j > 4$, implies (3.9).

k = 3: (3.8) for $\tau_1 = \ldots = \tau_4 = 0$, $\tau_5 = 1$, implies (3.9).

$i_k = 3,4$: $\xi_{i_k} = 3$

k = 0: $a_3 = a_4 = 0$ implies (3.9).

k = 1: (3) for $j = 3,4$ implies (3.9).

5. Methods of Order Eight.

Consider now the following ten systems of linear equations, each system having the same number of equations as unknowns.

(7) $\lambda_{21} = 1.$

(8) $$\sum_{j=1}^{2} \lambda_{3j} \mu_j^{\tau} = \frac{\mu_3^{\tau+1}}{\tau+1}, \quad \tau = 0, 1,$$

$$\left[\text{define } a_{31} = \frac{\sum_{j=1}^{2} \lambda_{3j} \mu_j^2 - \mu_3^{2/3}}{\lambda_{32} \mu_2^2}\right].$$

(9) $$\sum_{j=1}^{3} \lambda_{4j} \mu_j^{\tau} = \frac{\mu_4^{\tau+1}}{\tau+1}, \quad \tau = 0, 1,$$

$$\sum_{j=1}^{3} \lambda_{4j} \mu_j^2 - \frac{\lambda_{42} \mu_2^2}{a_{31}} = \frac{\mu_4^2}{3}.$$

(10) $$\sum_{j=1}^{4} \lambda_{5j} \mu_j^{\tau} = \frac{\mu_5^{\tau+1}}{\tau+1}, \quad \tau = 0, 1, 2,$$

$(\lambda_{52} = 0)$

$$\left[\text{define } a_{41} = \frac{\sum_{j=3}^{4} \lambda_{5j} \mu_j^3 - \mu_5^{3/4}}{\sum_{j=3}^{4} \lambda_{5j} \mu_j \lambda_{j2} \mu_2^2}\right].$$

(11) $$\sum_{j=1}^{5} \lambda_{6j} \mu_j^{\tau} = \frac{\mu_6^{\tau+1}}{\tau+1}, \quad \tau = 0, 1, 2,$$

$$\sum_{j=1}^{5} \lambda_{6j} \mu_j^3 - \sum_{j=1}^{5} \frac{\lambda_{6j} \mu_j \lambda_{j2} \mu_2^2}{a_{41}} = \frac{\mu_6^3}{4},$$

$(\lambda_{62} = 0).$

$$(12)\quad \sum_{j=1}^{6} \lambda_{7j}\,\mu_j^{\tau} = \frac{\mu_7^{\tau}}{\tau+1}\,, \qquad \tau = 0(1)2,$$

$$\sum_{j=1}^{6} \lambda_{7j}\,\mu_j^{3} = \sum_{j=1}^{6} \frac{\lambda_{7j}\,\mu_j\,\lambda_{j2}\,\mu_2^{2}}{a_{41}} = \frac{\mu_7^{3}}{4}\,,$$

$$\lambda_{73}\left[\lambda_{54}((\mu_6-\mu_7)\sum_{j=3}^{4}\lambda_{6j}\,\mu_j\,\lambda_{j2}) - \lambda_{64}((\mu_5-\mu_7)\sum_{j=3}^{4}\lambda_{5j}\,\mu_j\,\lambda_{j2})\right]$$

$$- \lambda_{74}\left[\lambda_{53}((\mu_6-\mu_7)\sum_{j=3}^{4}\lambda_{6j}\,\mu_j\,\lambda_{j2}) - \lambda_{63}((\mu_5-\mu_7)\sum_{j=3}^{4}\lambda_{5j}\,\mu_j\,\lambda_{j2})\right] = 0,$$

$$(\lambda_{72} = 0)\,.$$

$$(13)\quad \sum_{j=1}^{7} \lambda_{8j}\,\mu_j^{\tau} = \frac{\mu_8^{\tau}}{\tau+1}\,, \qquad \tau = 0(1)3,$$

$$(\lambda_{8j} = 0, \quad j = 2,3,4).$$

$$(14)\quad \sum_{j=1}^{8} \lambda_{9j}\,\mu_j^{\tau} = \frac{\mu_9^{\tau}}{\tau+1}\,, \qquad \tau = 0(1)3,$$

$$\sum_{i=8}^{11} a_i\,\mu_i(\mu_i-1)(\mu_i-\mu_{10})\sum_{j=5}^{i-1}\lambda_{ij}\,\mu_j\,\lambda_{j8} = a_8\left[\left(\tfrac{1}{4} - \tfrac{1}{3}\,\mu_8 + \frac{\mu_8^{4}}{12}\right)\right.$$

$$\left. - (1+\mu_{10})\left(\tfrac{1}{3} - \tfrac{1}{2}\,\mu_8 + \frac{\mu_8^{3}}{6}\right) + \mu_{10}\left(\tfrac{1}{2} - \mu_8 + \frac{\mu_8^{2}}{2}\right)\right]\,,$$

$$(\lambda_{9j} = 0, \quad j = 2,3,4)\,.$$

$$(15)\quad \sum_{j=1}^{9} \lambda_{10j}\,\mu_j^{\tau} = \frac{\mu_{10}^{\tau}}{\tau+1}\,, \qquad = 0(1)3,$$

$$\sum_{i=8}^{11} a_i\,\mu_i(\mu_i-1)\sum_{j=5}^{i-1}\lambda_{ij}\,\mu_j\,\lambda_{j8} = \frac{a_8}{6}\,(\mu_8-1)^{3}\,,$$

$$\sum_{i=8}^{11} a_i\,\mu_i(\mu_i-1)\sum_{j=5}^{i-1}\lambda_{ij}\,\mu_j^{2}\,\lambda_{j8} = \frac{a_8}{24}\,(3\mu_8+1)(\mu_8-1)^{3}\,,$$

$$(\lambda_{10j} = 0, \quad j = 2,3,4).$$

$$(16)\qquad \sum_{j=1}^{11} \lambda_{11j}\, \mu_j^{\tau} = \frac{1}{\tau+1}\,, \quad \tau = 0(1)3,$$

$$\sum_{i=8}^{11} a_i \mu_i \lambda_{ij} = a_j(1-\mu_j), \qquad j = 8, 9, 10,$$

$$(\lambda_{11j} = 0, \quad j = 2,3,4).$$

Theorem (3.4): If the choice of abscissae and weights is based on a 5-point Lobatto quadrature, then the remaining parameters are uniquely defined by conditions (7) to (16); further conditions (7) to (16) and conditions (1) to (6) are equivalent.

Proof: Each of the systems (7) to (16) has a unique solution if the matrix of coefficients is non-singular. For (7), (8), (10) and (13), the matrix has a van der Monde determinant, and is non-singular; for the other systems this is found to be the case by calculation for any ordering of the abscissae with $\mu_7 = \mu_8$ (see discussion below).

Next we observe that the $\{\lambda_{ij},\ j < i\}$ form a set of forty undetermined parameters after (1) is satisfied; further the sets (2) to (6) and (7) to (16) each have forty equations, and we show parameters satisfying the first set also satisfy the second set. Indeed, (7), (8), (10) and (13) are equivalent to (2) for the first abscissa μ_i in each of G_2, G_3, G_4, G_5 respectively. For $\mu_4 \in G_3$, (2), and (5) giving $v_{31} = a_{31} v_{32}$ imply (9); similarly for $\mu_6 \in G_4$, (2), and (5) giving $v_{41} = a_{41} v_{43}$ imply (11); (12) follows from these conditions for μ_7 together with (6). The first sets of equations of (14), (15) and (16) follow

from (2) with each of μ_9, μ_{10}, $\mu_{11} \in G_5$ respectively. The three remaining equations of (14) and (15) are derived from (4) for $\tau_1 + \tau_2 \leq 2$. (The proof of Theorem (3.3) implies (4) for $\tau_1 = 0$). The remaining equations of (16) are equivalent to (3).

Using similar arguments the reverse implications may be proved, and the result follows.

If the abscissae $\{\mu_i\}$ are chosen as described for a 5-point Lobatto quadrature, and further

$$\mu_7 = \mu_8 ,$$

then parameters calculated by the algorithm defined by conditions (7) to (16) satisfy the constraints which were not proved valid by Theorem (3.3). A general proof of this result has not been found. However, in the case that

$$\mu_2 = \mu_3 = \mu_6, \qquad \mu_4 = \mu_5 ,$$

it may be shown that the result follows on replacing conditions (4), (5) and (6) by conditions of the form

$$(3.12) \qquad \sum_{j=\ell+1}^{i-1} \lambda_{ij}\, \mu_j\, \underline{\lambda}_{j\ell} = \underline{\lambda}_{i\ell}\, (\mu_i - \mu_\ell)$$

for appropriate values of the indices. Indeed, the additional restrictions on the abscissae imply that (3.12) is trivially valid for the cases $(i = 8, \ \ell = 7)$, $(i = 5, \ \ell = 4)$ and $(i = 3, \ \ell = 2)$. Also the algorithm for calculating the parameters is simplified by (3.12).

Even more important, (3.12) is easily generalized for any value of p, the order of the method required. However, for $p > 8$,

the proof is not valid, and indeed, it appears that, for example, a tenth order method of this type does not exist. In attempting to extend Theorem (3.3) for higher values of p (for the above restrictions unnecessarily limit the ordering of the abscissae) conditions (1) to (5) generalize easily, but this is not true of (6). However, for p = 10, conditions corresponding to (6) may be derived, and on setting up the corresponding algorithm, it is found that no ordering of the abscissae will satisfy all the constraints.

6. Numerical Example

Here, one acceptable ordering of the abscissae is selected, and the parameters are calculated using single length arithmetic (Table (3.3)).

We consider the second order differential equation

$$(3.13) \qquad y'' = -y, \qquad y(0) = 0, \quad y'(0) = 1 ,$$

which has the solution

$$y(t) = \sin t ,$$

and the equivalent first order system

$$(3.14) \qquad \begin{aligned} y_1' &= y_2, & y_1(0) &= 0 , \\ y_2' &= -y_1, & y_2(0) &= 1 , \end{aligned}$$

which has the solution

$$y_1(t) = \sin t , \qquad y_2(t) = \cos t.$$

.0											
.1726731646	1										
.5	-.4478219619	1.4478219619									
.8273268354	.6577651537	-.7409902530	1.0832250993								
.5	.3992739873	0	.7545875386	-.1538615259							
.8273268354	.2242243882	0	.4289902366	.0787457947	.2680395805						
.1726731646	.7469967510	0	.4567051543	-.4949411950	-.0809956746	.3722349643					
.1726731646	.4136634177	0	0	0	-.0687349648	.0115951196	.6434764275				
.5	.0625	0	0	0	.3055555556	-.0181739222	-1.2650923214	1.9152106880			
.8273268354	.0863365823	0	0	0	-.7710753754	.1343013503	2.4549948730	-2.1888122148	1.2842547846		
1.0	0	0	0	0	2.4513804324	-.5512205631	-7.1649515531	7.5538404420	-2.2291582101	.9401094520	
	.05	0	0	0	0	0	0	.2722222222	.3555555556	.2722222222	.05

TABLE (3.3): An explicit Runge-Kutta method of order 8.

Using the recursion (3.7) to generate appropriate methods for a second order equation, these problems are solved by the methods of order four, six, and eight, given in Tables (3.2), (3.1) and (3.3) respectively with the steplength $h = 5$. The errors for the problems are given in Tables (3.4) and (3.5) respectively (with exponent to base 10 in brackets).

For each of these methods the errors for the two problems are comparable, and with an increase in order, there is the expected increase in accuracy.

x	Order	$\hat{y}-y$	$\hat{y}'-y'$
.5	4	-2.6 (-4)	2.2 (-5)
	6	1.0 (-5)	2.1 (-6)
	8	-6.0 (-8)	6.1 (-9)
20	4	-7.7 (-3)	6.9 (-3)
	6	-1.7 (-4)	1.1 (-4)
	8	-1.0 (-6)	1.0 (-6)
32	4	-1.6 (-2)	2.6 (-3)
	6	-3.4 (-4)	-3.4 (-5)
	8	-2.2 (-6)	5.4 (-7)

TABLE (3.4): Accummulated error in numerical solution of (3.13), $h = .5$.

x	Order	$\hat{y}_1-y_1$	$\hat{y}_2-y_2$
.5	4	-2.6 (-4)	2.2 (-5)
	6	2.0 (-7)	-9.7 (-8)
	8	-7.2 (-8)	-6.8 (-9)
20	4	-7.7 (-3)	6.9 (-3)
	6	4.1 (-6)	-8.0 (-6)
	8	-2.5 (-6)	1.5 (-6)
30	4	-1.6 (-2)	2.6 (-3)
	6	1.2 (-5)	-7.3 (-6)
	8	-4.6 (-6)	-4.8 (-8)

TABLE 3.5: Accumulated error in numerical solution of (3.14), $h = .5$.

Table (3.6) shows that more arithmetic operations are required per step when treating the problem as a second order differential equation. As the accuracies obtained are comparable

Order of Method	Number of arithmetic operations		Number of function evaluations
	(3.13)	(3.14)	
4	36	32	4
6	94	88	7
8	186	176	11

TABLE 3.6: Work required in one step.

in each case, it appears better to treat a problem as a first order system, in general.

CHAPTER IV

IMPLICIT RUNGE-KUTTA METHODS

1. Introduction

In an explicit method the restriction

$$\lambda_{rij}^{[\nu]} = 0 , \qquad j \geq i ,$$

is imposed; if this restriction is not imposed, methods of a given order requiring fewer stages may be obtained. However, such methods are implicit, and in general lead to algebraic equations which must be solved iteratively. Butcher [8] introduced such methods for a first order system of differential equations, and showed that some s-stage methods are of order 2s. Cooper [16] developed corresponding methods for a more general system and showed that a class of s-stage methods are of order* at least s. As the methods given by Cooper reduce to those of Butcher in the case of a first order system, it seemed reasonable to expect the more general methods were of order 2s - k for some integer k. Indeed, assuming that certain ordinary and partial derivatives exist for a system of differential equations, two types of s-stage methods are of order 2s. More generally there exist s-stage methods of order p for any $p \leq 2s$. (These results form the content of an article [39] included in the Appendix.)

Although the algebra is complicated, the principles are simple

---

* The definition of order here differs from that used by Cooper [16] and the author [39].

and we commence with an outline. Several of the arguments used here are similar to those of the previous chapter. The error terms, corresponding to those for explicit methods, are given by implicit equations, and may be expressed by a matrix equation. Provided that a sufficiently small stepsize is chosen, the matrix may be inverted, and expressions for these errors are derived in which all terms of order less than or equal to p are explicit. Sufficient conditions for a method to be of order p are thereby determined, givine (4.9). In Section 4 the order of one particular type of method is established. A lemma is first proved using a partial fraction expansion, equations (4.4), and the fact that a non-singular homogeneous equation has only the zero solution. The main result is then established using a double induction to show that the sufficient conditions (4.9) are satisfied.

2. Methods and Error Expressions

It is convenient to use the notation of Chapter III, and assuming

$$\hat{y}_r^{(n_r-\nu)}(x) = y_r^{(n_r-\nu)}(x) + O(h^{p+1}) ,$$

an s-stage implicit Runge-Kutta method may be written

$$\hat{y}_r^{(n_r-\nu)}(x_i) = \sum_{\tau=0}^{\nu-1} \frac{(\mu_i h)^\tau}{\tau!} \hat{y}_r^{(n_r-\nu+\tau)}(x) + \frac{(\mu_i h)^\nu}{\nu!} \sum_{j=1}^{s} \lambda_{rij}^{[\nu]} \hat{y}_r^{(n_r)}(x_j) ,$$

$$\hat{y}_r^{(n_r)}(x_i) = f_r(x_i;\ \{\hat{y}_\rho^{(n_\rho-m)}(x_i)\}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1 .$$

In particular we write $\alpha_{rj}^{[\nu]}$ for $\lambda_{r\,s+1\,j}^{[\nu]}$, and $\hat{\underline{y}}(x_{s+1})$ provides the required approximations.

These equations define a class of methods, subclasses of which are defined by constraints on the parameters. Here we are concerned with two subclasses, both satisfying

$$(4.1) \qquad \sum_{i=1}^{s} \alpha_{ri}^{[1]} \mu_i^{\tau} = \frac{1}{\tau+1}, \qquad \tau = 0(1)p'-1,$$

$$(4.1') \qquad \alpha_{ri}^{[\nu]} = \frac{\nu}{\nu-1}(1-\mu_i)\alpha_{ri}^{[\nu-1]}, \qquad \nu = 2(1)n_r,$$

$$r = 1(1)q, \quad i = 1(1)s,$$

where the order p of a method depends on p'.

The first subclass is then defined by

(i) $$(4.2) \qquad \sum_{j=1}^{s} \lambda_{rij}^{[1]} \mu_j^{\tau} = \frac{\mu_i^{\tau}}{\tau+1}, \qquad \tau = 0(1)s-1,$$

$$(4.2') \qquad \mu_i \lambda_{rij}^{[\nu]} = \frac{\nu}{\nu-1}(\mu_i-\mu_j)\lambda_{rij}^{[\nu-1]}, \qquad \nu = 2(1)n_r,$$

$$r = 1(1)q, \quad i = 1(1)s, \quad j = 1(1)s.$$

For the second subclass

(ii) $$(4.3) \qquad \sum_{j=1}^{s} \lambda_{rij}^{[\nu]} \mu_j^{\tau} = \mu_i^{\tau} \binom{\tau+\nu}{\tau}^{-1}$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s, \quad \tau = 0(1)s-1,$$

and in both cases the abscissae are chosen distinct. These are special cases of the methods considered by Cooper [16], and thus the order of a method satisfying (i) or (ii) is $p \geq \min\{p', s\}$.

Indeed, the parameters may be defined so that (4.2) or (4.3) is valid. For, in either case, each value of r, i and ν gives a system of s equations in s unknowns; for each system the matrix of coefficients is non-singular, and thus the system has a unique solution. By a similar argument $p' = s$ is possible for (4.1); thus some methods of the subclasses are of order s.

Further, similar results are possible for methods for systems of arbitrary orders. Such results are implied by the conclusions of a lemma.

Lemma (4.1): If the parameters are defined so that (4.1) and (4.1') are valid, then

$$(4.4)\qquad \sum_{i=1}^{s} a_{ri}^{[\gamma]} \mu_i^{\tau} = \binom{\tau+\gamma}{\tau}^{-1},$$

$$r = 1(1)q, \quad \gamma = 2(1)n_r, \quad \tau = 0(1)p' - \gamma.$$

Proof: The result for $\gamma = 1$ is equivalent to (4.1) and, we assume (4.4) is valid for $\gamma = 1(1)m$. Then for $\gamma = m+1$ we have

$$\sum_{i=1}^{s} a_{ri}^{[m+1]} \mu_i^{\tau} = \frac{m+1}{m} \sum_{i=1}^{s} a_{ri}^{[m]} (1-\mu_i)\mu_i^{\tau}$$

$$= \frac{m+1}{m} \left[\binom{\tau+m}{\tau}^{-1} - \binom{\tau+m+1}{\tau+1}^{-1} \right]$$

$$= \binom{\tau+m+1}{\tau}^{-1}$$

for $\tau+1 < p' - m+1$, or $\tau < p' - (m+1)+1$. Thus by induction on γ (4.4) is valid for $\tau < p' - \gamma+1$.

In a similar way, (4.2) and (4.2') imply

$$(4.5)\qquad \sum_{j=1}^{s} \lambda_{rij}^{[\nu]}\, \mu_j^{\tau} = \mu_i^{\tau} \binom{\tau+\nu}{\tau}^{-1},$$

$$r = 1(1)q, \quad \nu = 2(1)n_r, \quad i = 1(1)s, \quad \tau = 0(1)s-\nu .$$

By definition, a method is of order p if

$$\hat{y}_r^{(n_r-\nu)}(x+h) - y_r^{(n_r-\nu)}(x+h) = O(h^{p+1}), \quad r = 1(1)q, \quad = 1(1)n_r.$$

We wish to show that for certain choices of the abscissae $p > s$, and even $p = 2s$. Indeed, if the parameters are chosen to be zeros of the Legendre polynomial $P_s(2\mu-1)$, Butcher [8] has shown for first order systems that $p = 2s$. If the abscissae are chosen in this way, and $a_{ri}^{[\nu]}$ are the weights for the associated Gaussian quadrature on $(0, 1)$, then

$$\sum_{i=1}^{s} a_{ri}^{[1]}\, \mu_i^{\tau} = \int_0^1 \mu^{\tau}\, d\mu = \binom{\tau+1}{\tau}^{-1}, \quad \tau = 0(1)2s-1,$$

and thus (4.1) is valid for $p' = 2s$. For his proof Butcher requires also that (4.2) is valid, and as shown above the parameters may be chosen so that this is the case. Now the lemma implies that (4.4) is also valid for $\tau = 0(1)2s-\nu$ in this case.

By the proof below a method for which the parameters are chosen as for Gaussian quadratures is shown to be of order $p = 2s$; this is the maximum attainable order. By not requiring $p = 2s$, certain advantages may be gained. If the abscissae are

chosen differently, but still distinct, methods of order $2s-k$ for $k \geq 1$ may be obtained. For example, in Radau quadrature methods (Butcher [9]), which can be extended for systems of differential equations of arbitrary orders by (4.1') and (4.2'), one abscissa is chosen to be an endpoint of the interval $(0, 1)$, and $k = 1$. For methods based on Lobatto quadrature, with both endpoints abscissae, $k = 2$. We wish to be able to consider any of these choices, and thus continue to use the index p' where appropriate.

We proceed now to obtain error expressions. Following an argument similar to that for explicit Runge-Kutta methods, these expressions enable us to establish, in terms of parameter constraints, sufficient conditions for a method to be of order p.

We begin with errors* defined by

$$\psi_{ri}^{[\nu]} = \sum_{\tau=0}^{\nu-1} \frac{(\mu_i h)^{\tau}}{\tau!} y_r^{(n_r-\nu+\tau)}(x_i) + \frac{(\mu_i h)^{\nu}}{\nu!} \sum_{j=1}^{s} \lambda_{rij}^{[\nu]} y_r^{(n_r)}(x_j) - y_r^{(n_r-\nu)}(x_i),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1 ;$$

Taylor series expansions give

---

* Again $\psi_{r\,s+1}^{[\nu]}$ are the truncation errors and $\varepsilon_{r\,s+1}^{[\nu]}$ are the accumulated errors for a method.

$$(4.6)\quad \psi_{ri}^{[\nu]} = \sum_{\tau=0}^{p-\nu} \frac{\mu_i^{\nu} h^{\tau+\nu}}{(\tau+\nu)!} \left[\binom{\tau+\nu}{\tau} \sum_{j=1}^{s} \lambda_{rij}^{[\nu]} \mu_j^{\tau} - \mu_i^{\tau} \right] y_r^{(n_r+\tau)}(x) + O(h^{p+1}) .$$

Thus for a method from subclass (i), $\psi_{ri}^{[\nu]} = O(h^{s+1})$, and from subclass (ii) $\psi_{ri}^{[\nu]} = O(h^{s+\nu})$, for all values of the indices.

As before define

$$\varepsilon_{ri}^{[\nu]} = \hat{y}_r^{(n_r-\nu)}(x_i) - y_r^{(n_r-\nu)}(x_i) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)s+1,$$

and

$$\varepsilon_{ri} = f_r(x_i; \{\hat{y}_\rho^{(n_\rho-m)}(x_i)\}) - f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i)\})$$

$$r = 1(1)q, \quad i = 1(1)s .$$

These definitions imply that

$$\varepsilon_{ri} = f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i) + \psi_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} \varepsilon_{\rho j} + O(h^{p+1})\})$$

$$- f_r(x_i; \{y_\rho^{(n_\rho-m)}(x_i)\}) ,$$

$$r = 1(1)q, \quad i = 1(1)s,$$

and if, for example, the functions $f_r(t; \underline{y}(t))$ satisfy a Lipschitz condition with constant L

$$\left|\varepsilon_{ri}\right| \leq L \max_{\rho,m} \left| \psi_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} \varepsilon_{\rho j} + O(h^{p+1}) \right| ,$$

whence for $|h|$ sufficiently small

$$\varepsilon_{ri} = O(h^{s+1}) .$$

This result is used in the expansion (4.7) of the next section.

3. Parameter Constraints

Assume the existence of first partial derivatives

$$f_r(x_i)_{\rho m} = \left[\frac{\partial f_r(t; \{z_\eta^{[\nu]}\})}{\partial z_\rho^{[m]}} \right]_{(x_i; \{y_\eta^{(n_\eta - \nu)}(x_i)\})}$$

and of all total derivatives with respect to t thereof. A truncated Taylor expansion of ε_{ri} gives

$$(4.7) \quad \varepsilon_{ri} = \sum_{\rho=1}^{q} \sum_{m=1}^{n_\rho} \left[\psi_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} \varepsilon_{\rho j} \right] f_r(x_i)_{\rho m} + \phi_{ri} + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad i = 1(1)s,$$

where

$$\phi_{ri} = O(h^{2(s+1)}) .$$

As $p \leq 2s$, it is convenient to write the final terms of (4.7) as ϕ'_{ri} with $\phi'_{ri} = O(h^{p+1})$. In matrix notation (4.7) becomes

$$(4.8) \quad (I - Z)\varepsilon = \theta + \phi'$$

where I is the identity matrix, Z is an sq x sq matrix, $Z = \{z_{ri,\rho j}\}$ defined by

$$(4.8') \quad z_{ri,\rho j} = \sum_{m=1}^{n_\rho} \frac{(\mu_i h)^m}{m!} \lambda^{[m]}_{\rho ij} f_r(x_i)_{\rho m} ,$$

and ε, θ, ϕ' are sq-dimensional column vectors defined, for example, by the transpose

$$\varepsilon^T = (\varepsilon_{11}, \ldots, \varepsilon_{1s}; \varepsilon_{q1}, \ldots, \varepsilon_{qs}) ,$$

with

$$(4.8'') \quad \theta_{ri} = \sum_{\rho=1}^{q} \sum_{m=1}^{n_\rho} \psi^{[m]}_{\rho i} f_r(x_i)_{\rho m} ,$$

and for methods of subclass (i) or (ii), $\theta_{ri} = O(h^{s+1})$. As $m \geq 1$ in (4.8'), for all sufficiently small $|h|$ the eigenvalues of Z are less than unity, and there exists a power series expansion for $(I - Z)^{-1}$. Thus (4.8) may be written

$$(4.9) \quad \varepsilon = (I + Z + \ldots + Z^{p-s-1-\nu})\theta + O(h^{p+1-\nu}) .$$

Lemma (4.2): For a method to be of order p, it is sufficient that

$$(4.10) \quad \sum_{i_0=1}^{s} \alpha^{[m_0]}_{r_0 i_0} \mu_{i_0}^{m_1+\tau_1} \sum_{i_1=1}^{s} \lambda^{[m_1]}_{r_1 i_0 i_1} \mu_{i_1}^{m_2+\tau_2} \cdots$$
$$\cdots \mu_{i_{k-1}}^{m_k+\tau_k} \left[\binom{\tau_{k+1}+m_k}{\tau_{k+1}} \sum_{i_k=1}^{s} \lambda^{[m_k]}_{r_k i_{k-1} i_k} \mu_{i_k}^{\tau_{k+1}} - \mu_{i_{k=1}}^{\tau_{k+1}} \right] = 0$$

for non-negative integral k, $k \le p-s-\gamma$, with

$$\tau_1 + \ldots + \tau_{k+1} + m_0 + \ldots + m_k \le p, \quad \tau_j \ge 0, \quad m_j > 0.$$

Proof: By definition, a method is of order p if

$$\hat{y}_r^{(n_r-\gamma)}(x+h) - y_r^{(n_r-\gamma)}(x+h) = O(h^{p+1}) .$$

But the left side is equal to

$$\frac{h^\gamma}{\gamma!} \sum_{i=1}^{s} a_{ri}^{[\gamma]} \varepsilon_{ri} + \sum_{\tau=0}^{p-\gamma} \frac{h^{\tau+\gamma}}{(\tau+\gamma)!} \left[\binom{\tau+\gamma}{\tau} \sum_{i=1}^{s} a_{ri}^{[\gamma]} \mu_i^{\tau} - 1 \right] y_r^{(n_r+\tau)}(x) + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad \gamma = 1(1)n_r .$$

By (4.10) for $k = 0$ (which is equivalent to (4.1) and (4.4)), the second term of the right side is zero. To show that the first term is $O(h^{p+1})$, it is sufficient that

$$a_{r_o}^{[\gamma]} Z^{k-1} \Theta = O(h^{p+1-\gamma}),$$

$$r = 1(1)q, \quad \gamma = 1(1)n_r, \quad k = 1(1)p-s-\gamma ,$$

where

$$a_{r_o}^{[\gamma]} = (0, \ldots, 0; \ldots ; a_{r_o 1}^{[\gamma]}, \ldots, a_{r_o s}^{[\gamma]}; \ldots; 0, \ldots, 0) .$$

By an induction on k, the general element of Z^{k-1}

$$Z^{k-1} = \left\{ z_{r_o i_o}^{[k-1]}, \; r_{k-1} i_{k-1} \right\} , \quad k \ge 2 ,$$

is given by

$$z^{[k-1]}_{r_o i_o, r_{k-1} i_{k-1}} = \sum_{r_1=1}^{q} \sum_{i_1=1}^{s} \cdots \sum_{r_{k-2}=1}^{q} \sum_{i_{k-1}=1}^{s} z_{r_o i_o, r_1 i_1} \cdots$$

$$\cdots z_{r_{k-2}\, i_{k-2}, r_{k-1} i_{k-1}} ,$$

and z^o is the identity matrix. Thus

$$(4.11)\quad a_{r_o} z^{k-1} \theta = \sum_{i_o=1}^{s} \cdots \sum_{r_{k-1}=1}^{q} \sum_{i_{k-1}=1}^{s} a_{r_o i_o} z_{r_o i_o, r_1 i_1} \cdots$$

$$\cdots z_{r_{k-2} i_{k-2}, r_{k-1} i_{k-1}} \theta_{r_{k-1} i_{k-1}} .$$

The Taylor expansion

$$f_r(x_i)_{\rho m} = \sum_{\tau=0}^{p} \frac{(\mu_i h)^\tau}{\tau!} f_r^{(\tau)}(x)_{\rho m} + O(h^{p+1})$$

gives (4.8') as

$$z_{ri,\rho j} = \sum_{m=1}^{n_\rho} \frac{(\mu_i h)^m}{m!} \lambda^{[m]}_{\rho i j} \left\{ \sum_{\tau=0}^{p-m-\nu} \frac{(\mu_i h)^\tau}{\tau!} f_r^{(\tau)}(x)_{\rho m} \right\} + O(h^{p+1-\nu}),$$

and (4.8") with (4.6) as

$$\theta_{ri} = \sum_{\rho=1}^{q} \sum_{m=1}^{n_\rho} \left\{ \sum_{\tau=0}^{p-m-\nu} \frac{\mu_i^m h^{\tau+m}}{(\tau+m)!} \left[\binom{\tau+m}{\tau} \sum_{j=1}^{s} \lambda^{[m]}_{rij} \mu_j^\tau - \mu_i^\tau \right] y_r^{(n_r+\tau)}(x) \right\} \times$$

$$\left\{ \sum_{\sigma=0}^{p-1-\tau-m-\nu} \frac{(\mu_i h)^\sigma}{\sigma!} f_r^{(\sigma)}(x)_{\rho m} \right\} + O(h^{p+1-\nu}) .$$

Substituting these expressions in (4.11), it follows that conditions (4.10) are sufficient to imply the method is of order p.

Except for the range of summation for the indices $\{i_\ell, \ell = 0(1)k\}$, conditions (3.5) and (4.10) are identical. Further, for methods of subclass (i), (4.1') and (4.2') are equivalent to (3.7), and thus the conclusion of Theorem (3.2) is valid for such methods. Indeed, methods of this subclass are of order $p = p'$ as now follows from the proofs given by Butcher [8,9]; in particular, if the abscissae are zeros of the Legendre polynomial $P_s(2\mu-1)$, then $p = 2s$. As the proof of this result is more involved for methods of subclass (ii) (and further, the complete result for (i) follows by a similar proof), this case is considered further.

4. The Order of Certain Methods

Here, the parameters for a method of subclass (ii) are shown to satisfy (4.10) for $p = p'$. For simplicity we assume that

$$\alpha_{ri}^{[\nu]} = \alpha_i^{[\nu]}, \quad \lambda_{rij}^{[\nu]} = \lambda_{ij}^{[\nu]}, \quad n_r = n.$$

(Actually this assumes that each equation is treated in the same way, and restricts only the class of methods examined).

The basic result is presented in a lemma.

Lemma (4.3): For parameters of a method of subclass (ii),

$$(4.12)\qquad \sum_{i=1}^{s} \alpha_i^{[\gamma]} \mu_i^{m+\sigma} \left[\binom{m+\tau}{\tau} \sum_{j=1}^{s} \lambda_{ij}^{[m]} \mu_j^{\tau} - \mu_i^{\tau} \right] = 0 ,$$

$$\gamma, m = 1(1)n, \quad \sigma = 0(1)p'-\gamma-m, \quad \tau = 0(1)p'-\gamma-m-\sigma .$$

Proof: Condition (4.4) gives

$$\sum_{i=1}^{s} \alpha_i^{[\gamma]} \mu_i^{m+\sigma+\tau} = \binom{m+\sigma+\tau+\gamma}{\gamma}^{-1}, \qquad 0 \le m+\sigma+\tau \le p' - \gamma,$$

and we must show that

$$(4.13)\qquad \sum_{i=1}^{s} \alpha_i^{[\gamma]} \mu_i^{m+\sigma} \sum_{j=1}^{s} \lambda_{ij}^{[m]} \mu_j^{\tau} = \binom{m+\tau}{\tau}^{-1} \binom{m+\sigma+\tau+\gamma}{\gamma}^{-1} ,$$

$$0 \le \tau \le p'-\gamma-m-\sigma .$$

As the right side may be written with absent from the numerator and factors in the denominator distinct, it may be expanded in partial fractions

$$\frac{m!\ \tau!}{(m+\tau)!} \frac{\gamma!(m+\sigma+\tau)!}{(m+\sigma+\tau+\gamma)!} = \sum_{k=1}^{m} \frac{c_k}{\tau + k} + \sum_{\ell=1}^{\gamma} \frac{d_\ell}{m+\sigma+\tau+\ell}$$

where

$$c_k = \frac{m!}{(-k+1)\ldots(-1)(1)\ldots(-k+m)} \quad \frac{\gamma!}{(m+\sigma-k+1)\ldots(m+\sigma-k+\gamma)}$$

and

$$d_\ell = \frac{m!}{(-m-\sigma-\ell+1)\ldots(-m-\sigma-\ell+m)} \quad \frac{\gamma!}{(-\ell+1)\ldots(-1)(1)\ldots(-\ell+\gamma)}$$

are independent of τ. Then (4.7) implies that

$$(4.14)\quad \binom{m+\tau}{\tau}^{-1} \binom{m+\sigma+\tau+\gamma}{\gamma}^{-1} = \sum_{k=1}^{m} c_k \sum_{j=1}^{s} a_j^{[1]} \mu_j^{\tau+k-1} + \sum_{\ell=1}^{\gamma} d_\ell \sum_{j=1}^{s} a_j^{[1]} \mu_j^{m+\sigma+\tau+\ell-1}$$

for $0 < m + \sigma + \tau + \gamma \leq p'$. Now (4.3) implies (4.12), and thus (4.13), for $\tau = 0(1)s-1$. Further using (4.14), we obtain

$$\sum_{j=1}^{s} \mu_j^{\tau} \left[\sum_{i=1}^{s} a_i^{[1]} \mu_i^{m+\sigma} \lambda_{ij}^{[m]} - \sum_{k=1}^{m} c_k a_j^{[1]} \mu_j^{k-1} - \sum_{\ell=1}^{\gamma} d_\ell \mu_j^{m+\sigma+\ell-1} \right] = 0,$$

$$= 0(1)s-1,$$

where the expressions in square brackets are independent of τ. As the abscissae are distinct, the matrix of coefficients has a van der Monde determinant, and is therefore non-singular. Hence the expression in square brackets vanishes identically. Thus the value of τ is immaterial, and as (4.14) is valid for $m+\sigma+\tau+\gamma \leq p'$, the lemma is proved.

Theorem (4.1): For parameters satisfying the constraints (ii), (4.10) is valid for $p = p'$, thus giving methods of order p'.

Proof: The lemma establishes (4.10) for $k = 1$. From (4.4),

$$\sum_{i_o=1}^{s} a_{i_o}^{[m_o]} \mu_{i_o}^{m_1+\tau_1+\tau} = \binom{m_o + m_1 + \tau + \tau_1}{m_o}^{-1}$$

$$0 \leq m_1 + \tau_1 + \tau \leq p' - m_o \quad .$$

Now we assume that (4.10) is valid for all $k < \ell$, some positive integer, and further that

$$(4.15)\quad \sum_{i_0=1}^{s} \alpha_{i_0}^{[m_0]} \mu_{i_0}^{m_1+\tau_1} \sum_{i_1=1}^{s} \lambda_{i_0 i_1}^{[m_1]} \mu_{i_1}^{m_2+\tau_2} \cdots \sum_{i_{\ell-1}=1}^{s} \lambda_{i_{\ell-2} i_{\ell-1}}^{[m_{\ell-1}]} \mu_{i_{\ell-1}}^{\tau}$$

$$= \binom{m_{\ell-1}+\tau}{m_{\ell-1}}^{-1} \cdots \binom{m_0+\cdots m_{\ell-1}+\tau+\tau_1+\cdots+\tau_{\ell-1}}{m_0}^{-1},$$

$$\tau + (m_1+\cdots+m_{\ell-1}+\tau_1+\cdots+\tau_{\ell-1}) - 1 < p' - m_0,$$

and proceed to prove corresponding results for $k = \ell$.

As well as a result corresponding to (4.15) we must show that

$$(4.16)\quad \sum_{i_0=1}^{s} \alpha_{i_0}^{[m_0]} \mu_{i_0}^{m_1+\tau_1} \sum_{i_1=1}^{s} \lambda_{i_0 i_1}^{[m_1]} \mu_{i_1}^{m_2+\tau_2} \cdots \sum_{i_{\ell-1}=1}^{s} \lambda_{i_{\ell-2} i_{\ell-1}}^{[m_{\ell-1}]} \mu_{i_{\ell-1}}^{m_\ell+\tau_\ell} \sum_{i_\ell=1}^{s} \lambda_{i_{\ell-1} i_\ell}^{[m_\ell]} \mu_{i_\ell}^{\tau}$$

$$= \binom{m_\ell+\tau}{\tau}^{-1} \sum_{i_0=1}^{s} \alpha_{i_0}^{[m_0]} \mu_{i_0}^{m_1+\tau_1} \sum_{i_1=1}^{s} \lambda_{i_0 i_1}^{[m_1]} \mu_{i_1}^{m_2+\tau_2} \cdots \sum_{i_{\ell-1}=1}^{s} \lambda_{i_{\ell-2} i_{\ell-1}}^{[m_{\ell-1}]} \mu_{i_{\ell-1}}^{m_\ell+\tau_\ell+\tau},$$

$$\tau+(m_1+\cdots+m_\ell+\tau_1+\cdots+\tau_\ell) - 1 < p' - m_0,$$

whence (4.10) is valid for ℓ .

Setting $\tau = \tau + m_\ell + \tau_\ell$ in (4.15), the right side of (4.16) is equal to

$$(4.17) \quad \binom{m_\ell + \tau}{\tau}^{-1} \left[\binom{m_{\ell-1}+m_\ell+\tau_\ell+\tau}{m_{\ell-1}} \right]^{-1} \cdots \left[\binom{m_0+\cdots+m_\ell+\tau+\tau_1+\cdots+\tau_\ell}{m_0} \right]^{-1}$$

As in the lemma, (4.17) may be written with the numerator independent of τ, the denominator having distinct factors each linear in τ; thus it has a great partial fraction expansion with coefficients independent of τ. Now as (4.16) may be written in a form similar to (4.12), it is valid for $\tau = O(1)s-1$. As in the lemma the value of τ is immaterial, and thus (4.16) is established for other values of τ.

In the course of establishing (4.16), (4.15) was shown to be valid for $k = \ell$. Thus (4.10) follows by induction on k, and a method of subclass (ii) is of order p' .

It now appears reasonable to investigate the error of such methods of either subclass. Indeed, using the expression derived from the definition of order, leading error terms may be determined. However, numerical calculation of bounds for such terms is impractical. These terms are given in an appendix (reference [39]).

5. Numerical Example

Problems (3.13) and (3.14) solved by explicit methods are used as examples here. They are solved by corresponding implicit methods based on Lobatto quadratures - of the same orders. Parameters for the solution of (3.14) are given by Butcher [9], and additional parameters for (3.13) are generated by (4.1') and (4.2').

x	Order of Method	Error in y	Error in y'
.5	4	-2.6 (-4)	2.2 (-5)
	6	1.2 (-7)	8.3 (-8)
	8	-7.4 (-10)	1.6 (-10)
20	4	-2.9 (-3)	3.0 (-3)
	6	-5.1 (-6)	8.6 (-6)
	8	-8.3 (-9)	1.2 (-8)
32	4	-6.2 (-3)	1.8 (-3)
	6	-1.5 (-5)	6.6 (-6)
	8	-2.1 (-8)	9.7 (-9)

TABLE (4.1): Error in solution of (3.13) by implicit methods, $h = .5$.

Methods for problem (3.13) require fewer iterations than those for (3.14); indeed, for $s = 3$, the method for (3.13) is explicit. (As the results in Table (4.3) are for comparison

only, the average number of iterations was taken to the nearest integer.) Thus, if implicit methods are used, it may be more economical to treat a system of equations of arbitrary orders as such, than to reduce it to a first order system.

x	Order of Method	Error in y_1	Error in y_2
.5	4	6.2 (-5)	- 1.8(-5)
	6	9.6 (-8)	5.4 (-7)
	8	1.0 (-10)	- 8.9 (-11)
20	4	1.5 (-3)	- 2.1 (-3)
	6	2.2 (-6)	- 3.5 (-6)
	8	1.2 (-9)	- 5.2 (-9)
32	4	3.8 (-3)	- 1.5 (-3)
	6	5.9 (-6)	- 2.8 (-6)
	8	6.1 (-9)	- 5.9 (-9)

TABLE (4.2): Error in solution of (3.14) by implicit methods, $h = .5$.

A comparison of Tables (4.1) and (4.2) with Tables (3.4) and (3.5) respectively indicates that implicit methods of a particular order are more accurate than corresponding explicit methods. Further, implicit methods may be used for the solution of special problems. Indeed, this is the case for "stiff" differential equations (Cooper [18]) for which explicit methods are not adequate.

However, in general, implicit methods require more work per step than explicit methods. For the problems solved, this is evident from a comparison of Tables (4.3) and (3.6). (Here, iteration is stopped when each pair of successive iterates agrees to ten significant decimal digits.)

Order of Method	Average No. of Iterations	Average No. of Function Evaluations	Average No. of Arithmetic Operations
Problem (3.13)			
4	3	5	52
6	7	16	230
8	7	23	416
Problem (3.14)			
4	13	15	150
6	12	26	368
8	11	35	630

TABLE (4.3): Work per step required for implicit methods.

The work required for the iterative scheme may be reduced. Indeed, only several iterations are required if the initial approximations are good. Here, the initial approximations are obtained by finite-difference formulae; as the steplength is large, and the solutions oscillate, these are not very accurate. Better initial approximations are available from explicit methods

having the same abscissae. For example, consider the method of order 8 in Table (3.3). Three function evaluations give iterates with errors of $O(h^2)$; these are at least as accurate as the finite-difference approximations. Six function evaluations give iterates with errors of $O(h^4)$ and nine function evaluations give iterates with errors of $O(h^5)$. The iterative scheme could be started using any of these approximations. (Further, using the two methods simultaneously, an error estimate is available.) Still, additional iteration is usually required; hence we conclude that, when applicable, explicit methods are, in general, more economical than implicit methods.

CHAPTER V

QUADRATURES WITH WEIGHT FUNCTIONS

1. Introduction

In the numerical integration of a function, weight functions may be used to compensate for difficult behaviour provided that the nature of the singularity is known. Cooper [17] uses this technique to develop "quadrature" methods for a system of differential equations having singularities in one or more of the derivatives; further, certain classes of s-stage methods are of order p, with $p \geq s+1$.

These methods may be considered as implicit Runge-Kutta methods. Indeed, provided that the condition of existence of derivatives is satisfied, results of Chapter IV imply that for a unit weight function some methods are of order $2s$. Numerical results have indicated a similar result for certain variable weight functions. Here, such a result is given for a restricted class of problems using correspondingly restricted weight functions. More general results may be possible but, if so, a different analysis seems necessary. A numerical example* exhibits the increase in accuracy.

---

* A less significant increase in accuracy is exhibited by an implicit differential equation (7.6) treated by a related technique in Chapter VII.

2. Quadrature and Interpolation Methods

Here the methods are described. For an explicit differential problem, initial conditions are used to define

$$T_r^{[\nu]}(\mu h) = \sum_{\tau=0}^{\nu-1} \frac{(\mu h)^\tau}{\tau!} y_r^{(n_r-\nu+\tau)}(x), \qquad r = 1(1)q, \quad \nu = 1(1)n_r.$$

Consider a set of weight functions $\{\omega_r(t), \ r = 1(1)q\}$ such that

$$(5.1) \quad \omega_r(x+\mu h) = O((\mu h)^{\beta_r}), \quad \omega_r^{-1}(x+\mu h) = O((\mu h)^{-\beta_r}), \quad -1 < \beta_r < 1.$$

Then integration by parts gives

$$y_r^{(n_r-\nu)}(x+h) = T_r^{[\nu]}(h) + \frac{h^\nu}{(\nu-1)!} \int_0^1 g_r(x+\mu h)(1-\mu)^{\nu-1} \omega_r(x+\mu h)\, d\mu,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r,$$

where we assume the existence of p continuous derivatives of

$$g_r(t) = \omega_r^{-1}(t)\, y_r^{(n_r)}(t)$$

for some positive integer p, $p \geq s+\nu$, in a neighbourhood of $t = x$.

A set of distinct abscissae $\{\mu_i, \ i = 1(1)s\}$ defines points $x_i = x + \mu_i h$ for some steplength h, and an approximation to $\underline{y}(x+h)$ is given by

$$(5.2) \quad \tilde{y}_r^{(n_r-\nu)}(x+h) = T_r^{[\nu]}(h) + \frac{h^\nu}{\nu!} \sum_{i=1}^{s} a_{ri}^{[\nu]} g_r(x_i)$$

$$r = 1(1)q, \quad \nu = 1(1)n_r,$$

where the weights $\{a_{ri}^{[\nu]}\}$ are to be determined by some quadrature

rule. Indeed, we may interpolate either set of functions,

(i) $g_r(x + \mu h)$, $\quad r = 1(1)q$,

(ii) $(1 - \mu)^{\gamma-1} g_r(x + \mu h)$, $\quad r = 1(1)q, \ \gamma = 1(1)n_r$,

using the fundamental polynomials $l_i(\mu)$ of degree s-1 defined by

$$l_i(\mu) = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

For (i) this gives for appropriate indices

$$a_{ri}^{[\gamma]} = \gamma \int_0^1 l_i(\mu)(1-\mu)^{\gamma-1} \omega_r(x + \mu h)\, d\mu \ ,$$

and

$$y_r^{(n_r-\gamma)}(x+h) - \tilde{y}^{(n_r-\gamma)}(x+h)$$

$$= \frac{h^{\gamma}}{(\gamma-1)!} \int_0^1 \left\{ g_r(x+\mu h) - \sum_{i=1}^{s} l_i(\mu) g_r(x+\mu_i h) \right\} (1-\mu)^{\gamma-1} \omega_r(x+\mu h)\, d\mu$$

$$= O(h^{s+\gamma+\beta_r})$$

as the abscissae are distinct. Further, this error is $O(h^{2s+1+\beta})$ if only one weight function is used, and the abscissae are selected as the zeros of the orthogonal polynomial associated with this weight function on the interval $[0,1]$. A method based on these abscissae is of some advantage for problems in which each differential equation has the same singularity; this is examined in more detail later.

For (ii) we obtain for appropriate indices

$$\alpha_{ri}^{[\gamma]} = \gamma(1-\mu_i)^{\gamma-1} \int_0^1 l_i(\mu)\omega_r(x+\mu h)\,d\mu$$

and

$$y_r^{(n_r-\gamma)}(x+h) - \tilde{y}_r^{(n_r-\gamma)}(x+h) = \frac{h^\gamma}{(\gamma-1)!} \int_0^1 \Big\{ g_r(x+\mu h)(1-\mu)^{\gamma-1} - \sum_{i=1}^{s} l_i(\mu) g_r(x+\mu_i h)(1-\mu_i)^{\gamma-1} \Big\} \omega_r(x+\mu h)\,d\mu$$

$$= O(h^{s+1+\beta_r})$$

as the abscissae are distinct. Again this error is $O(h^{2s+1+\beta})$ if for a single weight function, the abscissae are selected as the zeros of the corresponding orthogonal polynomial.

To apply (5.2) values of $g_r(x_i)$ are required. These are not available; however approximations g_{ri} may be determined implicitly from the differential equations. First, integration by parts gives for appropriate indices

$$y_r^{(n_r-\gamma)}(x_i) = T_r^{[\gamma]}(\mu_i h) + \frac{h^\gamma}{(\gamma-1)!} \int_0^{\mu_i} (\mu_i-\mu)^{\gamma-1} g_r(x+\mu h)\omega_r(x+\mu h)\,d\mu ,$$

and again we consider two cases corresponding to the choices of interpolation above. We obtain

$$(5.3)\quad y_r^{(n_r-\gamma)}(x_i) = T_r^{[\gamma]}(\mu_i h) + \frac{(\mu_i h)^\gamma}{\gamma!} \sum_{j=1}^{s} \lambda_{rij}^{[\gamma]} g_r(x_j) + R_r^{[\gamma]}(\mu_i h)$$

where, in the first case

$$\lambda_{rij}^{[\gamma]} = \frac{\gamma}{\mu_i^{\gamma}} \int_0^{\mu_i} l_j(\mu)(\mu_i-\mu)^{\gamma-1}\, \omega_r(x+\mu h)\, d\mu \ ,$$

$$R_r^{[\gamma]}(\mu_i h) = \frac{h^{\gamma}}{(\gamma-1)!} \int_0^{\mu_i} \left\{ g_r(x+\mu h) - \sum_{j=1}^{s} l_j(\mu) g_r(x_j) \right\} (\mu_i-\mu)^{\gamma-1} \cdot \omega_r(x+\mu h)\, d\mu$$

$$= O(h^{s+\gamma+\beta_r}) \ ,$$

and in the second case

$$\lambda_{rij}^{[\gamma]} = \frac{\gamma}{\mu_i} \left(1 - \frac{\mu_j}{\mu_i}\right)^{\gamma-1} \int_0^{\mu_i} l_j(\mu)\omega_r(x+\mu h)\, d\mu \ ,$$

$$R_r(\mu_i h) = \frac{h^{\gamma}}{(\gamma-1)!} \int_0^{\mu_i} \left\{ g_r(x+\mu h)(\mu_i-\mu)^{\gamma-1} - \sum_{j=1}^{s} l_j(\mu)(\mu_i-\mu_j)^{\gamma-1} g_r(x_j) \right\} \omega_r(x+\mu h)\, d\mu$$

$$= O(h^{s+1+\beta_r}) \ .$$

A comparison of these two quadrature errors respectively with the constraints of subclasses (i) and (ii) of Chapter IV. leads to the conclusion that the parameters for those subclasses may be derived using these quadratures with a unit weight function.

Now substituting (5.3) into (1.2) gives

$$y_r^{(n_r)}(x_i) = f_r\left(x_i\, ; \left\{ T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]}\, g_\rho(x_j) + R_\rho^{[m]}(\mu_i h) \right\}\right) \ ,$$

$$r = 1(1)q, \quad i = 1(1)s \ ,$$

and approximationscorresponding to $g_r(x_j)$ are defined by

$$g_{ri} = \omega_r^{-1}(x_i) f_r\Big(x_i ; \Big\{ T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} \, g_{\rho j} \Big\}\Big);$$

approximations corresponding to $\hat{y}_r^{(n_r-\gamma)}(x+h)$ of (5.2) are denoted by $\hat{y}_r^{(n_r-\gamma)}(x+h)$.

Defining

$$\varepsilon_{ri} = g_{ri} - g_r(x_i)$$

gives

$$(5.4)\quad \varepsilon_{ri} = \omega_r^{-1}(x_i) f_r\Big(x_i; \Big\{ T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} (g_\rho(x_j) + \varepsilon_{\rho j}) \Big\}\Big)$$

$$- \omega_r^{-1}(x_i) f_r\Big(x_i; \Big\{ T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]} \, g_\rho(x_j) + R_\rho^{[m]}(\mu_i h) \Big\}\Big).$$

Assuming, for example, a Lipschitz condition, it may be shown by the proof given by Cooper [17] that at least

$$(5.5)\qquad \varepsilon_{ri} = O(h^{s+1}) .$$

This result is assumed in the development of the next section.

The definitions and results of this section are used in Chapter VII for the derivation of numerical methods for implicit differential equations.

3. Error Expressions

For methods of order $p > s+1$, in general, it is necessary that a single weight function be used. Thus restriction to a

single differential equation leads to certain simplifications. With appropriate modifications to the arguments, the results are valid for a system of differential equations using a single weight function.

Definitions of error and corresponding formulae are analogous to those in the corresponding section of the previous chapter. Again we assume the existence of all first and second partial derivatives, and define

$$f(x_i)_m = \left[\frac{\partial f(t;\ \{y^{[\nu]}\})}{\partial y^{[m]}}\right]_{(x_i;\ \{y^{(n-\nu)}(x_i)\})}$$

$$i = 1(1)s, \quad m = 1(1)n .$$

The class of problems is restricted by assuming that $f(x+\mu_i h)_m$ has a Taylor series expansion about x. It is not known how restrictive this condition is; however, it appears to be valid for many problems for which $g(x+\mu h)$ has a Taylor series expansion about x.

A Taylor series expansion of (5.4) gives

$$\varepsilon_i = \sum_{m=1}^{n} \left[\frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{ij}\, \varepsilon_j^{[m]} - R^{[m]}(\mu_i h)\right] f(x_i)_m\, \omega^{-1}(x_i) + \phi_i ,$$

$$i = 1(1)s,$$

where ϕ_i is a sum of products of squares of expressions in square brackets with second partial derivatives of the functions and inverse weight functions; further, using the result for the quadrature errors and (5.5),

$$\phi_i = O(h^{2s+2+\beta_r}), \qquad i = 1(1)s .$$

In matrix notation this becomes

$$(I - Z)\underline{\varepsilon} = \underline{\psi} + \underline{\phi}$$

where the s x s matrix $Z = \{z_{ij}\}$ is defined by

$$z_{ij} = \sum_{m=1}^{n} \frac{(\mu_i h)^m}{m!} \lambda_{ij}^{[m]} f(x_i)_m \omega^{-1}(x_i) ,$$

and $\underline{\varepsilon}$, $\underline{\psi}$, $\underline{\phi}$ are s-dimensional column vectors

$$\psi_i = - \sum_{m=1}^{n} R^{[m]}(\mu_i h) f(x_i)_m \omega^{-1}(x_i) = O(h^{s+1}) .$$

By definition a method is of order p if

$$\hat{y}^{(n-\nu)}(x+h) - y^{(n-\nu)}(x+h) = O(h^{p+1}) \quad \nu = 1(1)n.$$

If h is sufficiently small, $(I - Z)^{-1}$ has a power series expansion which gives

$$\underline{\varepsilon} = (I + Z + \ldots + Z^{p-s-1})\underline{\psi} + O(h^{p+1}) .$$

Further,

$$\begin{aligned} \hat{y}^{(n-\nu)}(x+h) - y^{(n-\nu)}(x+h) &= (\hat{y}^{(n-\nu)}(x+h) - \tilde{y}^{(n-\nu)}(x+h)) \\ &\quad + (\tilde{y}^{(n-\nu)}(x+h) - y^{(n-\nu)}(x+h)) \\ &= \frac{h^\nu}{\nu!} \sum_{i=1}^{s} \alpha_i^{[\nu]} \varepsilon_i + (\tilde{y}^{(n-\nu)}(x+h) - y^{(n-\nu)}(x+h)), \end{aligned}$$

where, for the quadratures discussed, the final term is $O(h^{2s+1+\beta})$. As the maximum attainable order is $p = 2s$, a method is of order

$p < 2s$ (or $p \leq 2s$ if $\beta \geq 0$) if

$$\frac{h^{\nu}}{\nu!} \sum_{i=1}^{s} \alpha_i^{[\nu]} \varepsilon_i = O(h^{p+1}) , \qquad \nu = 1(1)n .$$

Thus we must show that

$$(5.6) \quad \underline{a}^{[\nu]} Z^{\ell-1} \underline{\psi} = O(h^{p-\nu+1}) , \qquad \ell = 1(1)p-\nu-s .$$

4. The Order of Certain Methods

Here we consider a more general class of quadrature rules. Indeed, choose as abscissae the zeros of a polynomial of degree $s, \hat{P}(\mu)$, which is orthogonal on $[0,1]$, with respect to a single weight function, to every polynomial of degree $s-k$, $k > 0$. Then a method derived using the appropriate quadrature formulae of Section 2 is of order $2s-k$ for a differential equation having the properties previously discussed.

The result of the following lemma is used several times in the proof of Theorem (5.1).

Lemma (5.1): Choose the abscissae and parameters as above. If a function $d_\nu(t)$ has at least $(2s-\nu-k+2)$ continuous derivatives with respect to t on $[0,h]$, then

$$(5.7) \quad \nu \int_0^1 d_\nu(\mu h)(h-\mu h)^{\nu-1} \omega(x+\mu h) h \, d\mu = h^{\nu} \sum_{i=1}^{s} \alpha_i^{[\nu]} d_\nu(\mu_i h) + O(h^{2s+\beta-k+2})$$

Proof: This result is a consequence of the quadrature rules, and has already been used in Section 2. For completeness, a proof is given. Let $\pi_k(\mu)$ represent an arbitrary polynomial of degree

less than or equal to k.

Again there are two cases; for the first with $\tau < 2s-k-\gamma+2$,

$$\int_0^1 (\mu h)^{\tau} (h-\mu h)^{\gamma-1} \omega(x+\mu h)h\, d\mu - \frac{h^{\gamma}}{\gamma} \sum_{i=1}^{s} \alpha_i^{[\gamma]} (\mu_i h)$$

$$= h^{\tau+\gamma} \int_0^1 \left\{ \mu^{\tau} - \sum_{i=1}^{s} l_i(\mu)\mu_i^{\tau} \right\} (1-\mu)^{\gamma-1} \omega(x+\mu h)\, d\mu$$

$$= h^{\tau+\gamma} \int_0^1 \left\{ \hat{P}(\mu)\pi_{s-k-\gamma+1}(\mu) + \tilde{\pi}_{s-1}(\mu) \right\} (1-\mu)^{\gamma-1} \omega(x+\mu h)\, d\mu ,$$

and by orthogonality this becomes

$$h^{\tau+\gamma} \int_0^1 \tilde{\pi}_{s-1}(\mu)\, (1-\mu)^{\gamma-1} \omega(x+\mu h)\, d\mu$$

$$= h^{\tau+\gamma} \int_0^1 \sum_{j=1}^{s} l_j(\mu)\, \tilde{\pi}_{s-1}(\mu_j) (1-\mu)^{\gamma-1} \omega(x+\mu h)\, d\mu$$

$$= h^{\tau+\gamma} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \tilde{\pi}_{s-1}(\mu_j)$$

$$= h^{\tau+\gamma} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \left\{ \hat{P}_s(\mu_j)\, \pi_{s-k-\gamma+1}(\mu_j) + \tilde{\pi}_{s-1}(\mu_j) \right\}$$

$$= h^{\tau+\gamma} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \left\{ \mu_j^{\tau} - \sum_{i=1}^{s} l_i(\mu_j)\mu_i^{\tau} \right\} = 0 .$$

Now (5.1) implies that

$$\int_0^1 \omega(x+\mu h)\, d\mu = O(h^{\beta}) .$$

The result follows by expanding $d_{\gamma}(\mu h)$ and $d_{\gamma}(\mu_i h)$ in truncated Taylor series.

This argument may be appropriately modified to give the corresponding result in the second case.

The weight function is now further restricted to a class for which

$$(5.8)\quad \left[\int_0^{\hat{\mu}} (\mu h)^{\tau}\, \omega(x+\mu h)h\, d\mu\right] \omega^{-1}(x+\hat{\mu}h)\,, \qquad \tau = 0,1,\ldots,$$

has a Taylor series expansion about $\bar{x}$. (For example, the weight function $\omega(t) = t^{\alpha}$, $-1 > \alpha > 1$, satisfies (5.1) and (5.8).)

The basic result is now presented in a lemma.

Lemma (5.2): If abscissae and parameters are determined as above, then

$$h^{\nu} \sum_{i=1}^{s} a_i^{[\nu]} \psi_i = O(h^{2s+\beta-k+2})$$

for a problem satisfying the restrictions of Section 3.

Proof: As the proof is similar for both cases, only the first case is considered. To apply Lemma (5.1) we must establish the existence of certain derivatives.

(a) As $g(x+\mu h)$ has a Taylor series expansion we may write

$$\left[\int_0^{\hat{\mu}} \left\{g(x+\mu h) - \sum_{i=1}^{s} l_i(\mu) g(x+\mu_i h)\right\} (h\hat{\mu} - h\mu)^{m-1} \omega(x+\mu h)h\, d\mu \right.$$

$$\left. \omega^{-1}(x+\hat{\mu}h)\right] f(x+\hat{\mu}h)_m$$

$$= \left[\int_0^{\hat{\mu}} \left\{\sum_{\tau=s}^{\bar{p}} g^{(\tau)}(x) h^{\tau} (\mu^{\tau} - \sum_{i=1}^{s} l_i(\mu)\mu_i^{\tau})\right\} (h\hat{\mu}-h\mu)^{m-1} \omega(x+\mu h)h\, d\mu \right.$$

$$\left. \omega^{-1}(x+\hat{\mu}h)\right] f(x+\hat{\mu}h)_m + O(h^{\bar{p}+m+1}) \quad .$$

Using (5.8) the first term on the right side may be written as a continuous function of $t = \hat{\mu}h$ (and h) on $[0,h]$ for which all derivatives with respect to t exist for any choice of $\bar{p} \geq 0$.

Similarly, it follows that

$$\int_{\hat{\mu}}^{1} (h\mu - h\hat{\mu})^{m-1} f(x+\mu h)_m (h-h\mu)^{\gamma-1} h \, d\mu \left\{ \sum_{\tau=s}^{\bar{p}} g^{(\tau)}(x) h^{\tau} (\hat{\mu}^{\tau} - l_i(\hat{\mu}) \mu_i^{\tau}) \right\}$$

may be written as a continuous function of $t = \hat{\mu}h$ (and h) on $[0, h]$, for which all derivatives with respect to t exist.

Now (5.6) for $\ell=1$ gives

$$\frac{h^{\gamma}}{\gamma!} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \psi_j = -\frac{h^{\gamma}}{\gamma!} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \sum_{m=1}^{q} R^{[m]}(\mu_j h) f(x+\mu_j h)_m \omega^{-1}(x+\mu_j h) .$$

For each value of m we may write

$$\frac{h^{\gamma}}{\gamma!} \sum_{j=1}^{s} \alpha_j^{[\gamma]} R^{[m]}(\mu_j h) f(x+\mu_j h)_m \omega^{-1}(x+\mu_j h)$$

$$= \frac{h^{\gamma}}{\gamma!} \sum_{j=1}^{s} \alpha_j^{[\gamma]} \left[\frac{1}{(m-1)!} \int_0^{\mu_j} \left\{ g(x+\mu h) - \sum_{i=1}^{s} l_i(\mu) g(x+\mu_i h) \right\} (h\mu_j - h\mu)^{m-1} \right.$$

$$\left. \omega(x+\mu h) h \, d\mu \cdot f(x+\mu_j h)_m \, \omega^{-1}(x+\mu_j h) \right]$$

$$= \frac{h^{\gamma}}{\gamma!} \sum_{j=1}^{s} \frac{\alpha_j^{[\gamma]}}{(m-1)!} \left[\int_0^{\mu_j} \left\{ \sum_{\tau=s}^{2s-k-\gamma+1} g^{(\tau)}(x) h^{\tau} (\mu^{\tau} - \sum_{i=1}^{s} l_i(\mu) \mu_i^{\tau}) \right\} \right.$$

$$\left. (h\mu_j - h\mu)^{m-1} \omega(x+\mu h) h \, d\mu \cdot f(x+\mu_j h)_m \, \omega^{-1}(x+\mu_j h) \right] + O(h^{2s-k+m+2}).$$

The existence of continuous derivatives implies (5.7) may be applied giving

$$\frac{1}{(\nu-1)!}\int_0^1 \frac{1}{(m-1)!}\left[\int_0^{\hat{\mu}} \left\{\sum_{\tau=s}^{2s-k-\nu+1} g^{(\tau)}(x)h^{\tau}\left(\mu^{\tau}-\sum_{i=1}^{s} l_i(\mu)\mu_i^{\tau}\right)\right\}(h\hat{\mu}-h\mu)^{m-1}\right.$$

$$\left.\omega(x+\mu h)h\,d\mu\cdot\omega^{-1}(x+\hat{\mu}h)f(x+\hat{\mu}h)_m\right](h-h\hat{\mu})^{-1}\omega(x+\hat{\mu}h)h\,d\hat{\mu}$$

$$+\ O(h^{2s+\beta-k+2})\ +\ O(h^{2s-k+m+2})\ .$$

As the appropriate integrals exist, the order of integration may be reversed. Since $m \geq 1$, the order terms may be condensed giving

$$\frac{1}{(\nu-1)!}\int_0^1 \frac{1}{(m-1)!}\left[\int_{\mu}^1 (h\hat{\mu}-h\mu)^{m-1} f(x+\hat{\mu}h)_m(h-h\hat{\mu})^{\nu-1} h\,d\hat{\mu}\right]\cdot$$

$$\left\{\sum_{\tau=s}^{2s-k-\nu+1} g^{(\tau)}(x)h^{\tau}\left(\mu^{\tau}-\sum_{i=1}^{s} l_i(\mu)\mu_i^{\tau}\right)\right\}\omega(x+\mu h)h\,d\mu + O(h^{2s+\beta-k+2}).$$

Again (5.7) may be applied to give

$$\frac{1}{(\nu-1)!}\,\frac{h}{(m-1)!}\sum_{i=1}^{s} a_j^{[1]}\left[\int_{\mu_j}^1 (h\hat{\mu}-h\mu)^{m-1} f(x+\hat{\mu}h)_m(h-h\hat{\mu})^{\nu-1} h\,d\hat{\mu}\right]\cdot$$

$$\left\{\sum_{\tau=s}^{2s-k-\nu+1} g^{(\tau)}(x)h^{\tau}\left(\mu_j^{\tau}-\sum_{i=1}^{s} l_i(\mu_j)\mu_i^{\tau}\right)\right\} + O(h^{2s+\beta-k+2}).$$

As the term in curlicue brackets is identically zero, the result follows.

Theorem (5.1): For a method defined in Lemma (5.2), $p = 2s-k$ (and $p = 2s-k+1$ for $\beta \geq 0$) for problems

satisfying the restrictions previously discussed.

Proof: Again it is necessary to consider only the first case.

Here we show for any choice of indices $(\nu, m_\ell, \ldots, m_1)$ that the corresponding component of $\underline{a}^{[\nu]} Z^{\ell-1} \underline{\psi}$ satisfies

$$(5.9)\quad h^\nu \left[\underline{a}^{[\nu]} Z^{\ell-1} \underline{\psi}\right]_{(\nu, m_\ell, \ldots, m_1)} = O(h^{2s+\beta-k+2}),\ \ell = 1(1)s-k-\nu+1.$$

Define

$$G^{[1]}(\mu h) = \frac{1}{(m_1-1)!} \int_0^\mu \left\{ g(x+\mu_0 h) - \sum_{i_0=1}^{s} l_{i_0}(\mu_0) g(x+\mu_{i_0} h) \right\} \times$$

$$(h\mu - h\mu_0)^{m_1-1} \omega(x+\mu_0 h) h\, d\mu_0 \cdot \omega^{-1}(x+\mu h) f(x+\mu h)_{m_1},$$

$$G^{[\ell]}(\mu h) = \frac{1}{(m_\ell-1)!} \int_0^\mu \left\{ \sum_{i_{\ell-1}=1}^{s} l_{i_{\ell-1}}(\mu_{\ell-1})\, G^{[\ell-1]}(\mu_{i_{\ell-1}} h) \right\} \times$$

$$(h\mu - h\mu_{\ell-1})^{m_\ell-1} \omega(x+\mu_{\ell-1} h) h\, d\mu_{\ell-1} \cdot f(x+\mu h)_{m_\ell}\, \omega^{-1}(x+\mu h),$$

$$\ell = 2(1)s-k-\nu+1 .$$

Then by definitions of Sections 2 and 3, (5.9) is valid if

$$(5.10)\quad \frac{h^\nu}{\nu!} \sum_{i_\ell=1}^{s} a_{i_\ell}^{[\nu]}\, G^{[\ell]}(\mu_{i_\ell} h) = O(h^{2s+\beta-k+2}),\quad \ell = 1(1)s-k-\nu+1.$$

Define

$$\overline{G}^{[1]}(\mu h) = \frac{1}{(m_1-1)!} \int_0^\mu \left\{ \sum_{\tau=s}^{2s-k-\nu+1} g^{(\tau)}(x) h^\tau \left(\mu_0^\tau - \sum_{i_0=1}^{s} l_{i_0}(\mu_0) \mu_{i_0}^\tau\right) \right\} \times$$

$$(h\mu - h\mu_0)^{m_1-1} \omega(x+\mu_0 h) h\, d\mu_0 \cdot \omega^{-1}(x+\mu h) f(x+\mu h)_{m_1},$$

$$\overline{G}^{[\ell]}(\mu h) = \frac{1}{(m_\ell-1)!}\int_0^\mu \Big\{\overline{G}^{[\ell-1]}(\mu_{\ell-1}h) - \sum_{\tau=s}^{2s-k-\nu+1} \overline{G}^{[\ell-1](\tau)}(0)h^\tau(\mu_{\ell-1}^\tau - \sum_{i_{\ell-1}=1}^{s} 1_{i_{\ell-1}}(\mu_{\ell-1})\mu_{i_{\ell-1}}^\tau)\Big\}\times$$

$$(h\mu-h\mu_{\ell-1})^{m_\ell-1}\,\omega(x+\mu_{\ell-1}h)h\,d\mu_{\ell-1}\cdot f(x+\mu h)_{m_\ell}\omega^{-1}(x+\mu h),$$

$$\ell = 2(1)s-k-\nu+1 .$$

Then $\overline{G}^{[\ell]}(t)$ has all derivatives continuous on $[0,h]$, and

$$G^{[\ell]}(\mu h) = \overline{G}^{[\ell]}(\mu h) + O(h^{2s-k-\nu+m_\ell+2}), \quad \ell = 1(1)s-k-\nu+1.$$

Indeed, for $\ell = 1$, these results follow from the proof of the lemma. Assume the results are valid for $\ell-1 > 0$. Then for ℓ set

$$\delta(\mu,\mu_{\ell-1}) \equiv (h_\mu-h_{\mu_{\ell-1}})^{m_\ell-1}\,\omega(x+\mu_{\ell-1}h)h\,d\mu_{\ell-1}\cdot f(x+\mu h)_{m_\ell}\,\omega^{-1}(x+\mu h),$$

and this gives

$$G^{[\ell]}(\mu h) = \frac{1}{(m_\ell-1)!}\int_0^\mu \Big\{\sum_{i_{\ell-1}=1}^{s} 1_{i_{\ell-1}}(\mu_{\ell-1})\,G^{[\ell-1]}(\mu_{i_{\ell-1}}h)\Big\}\times\delta(\mu,\mu_{\ell-1})$$

$$= \frac{1}{(m_\ell-1)!}\int_0^\mu \Big\{\sum_{i_{\ell-1}=1}^{s} 1_{i_{\ell-1}}(\mu_{\ell-1})(\overline{G}^{[\ell-1]}(\mu_{i_{\ell-1}}h) + O(h^{2s-k-\nu+2}))\Big\} \times\delta(\mu,\mu_{\ell-1})$$

$$= \frac{1}{(m_\ell-1)!}\int_0^\mu \Big\{\overline{G}^{[\ell-1]}(\mu_{\ell-1}h) - \sum_{\tau=s}^{2s-k-\nu+1} \overline{G}^{[\ell-1](\tau)}(0)\,h^\tau(\mu_{\ell-1}^\tau - \sum_{i_{\ell-1}=1}^{s} 1_{i_{\ell-1}}(\mu_{\ell-1})\mu_{i_{\ell-1}}^\tau)\Big\}\times\delta(\mu,\mu_{\ell-1}) + O(h^{2s-k-\nu+m_\ell+2})$$

since all derivatives for $\bar{G}^{[\ell-1]}(\mu h)$ exist. As the final expression except for the order term is $\bar{G}^{[\ell]}(\mu h)$, the second result is immediate. The existence of derivatives for $\bar{G}^{[\ell]}(t)$ follows as in Lemma (5.2). Thus by induction on ℓ, these results are valid for $\ell = 1(1)s-\nu-k+1$. Further (5.10) is valid if

$$(5.11)\quad \frac{h^{\nu}}{\nu!} \sum_{i_\ell=1}^{s} a_{i_\ell}^{[\nu]} \bar{G}^{[\ell]}(\mu_{i_\ell} h) = O(h^{2s+\beta-k+2}), \quad \ell = 1(1)s-k- +1,$$

since $m_\ell > 0$. Lemma (5.2) implies (5.11) for $\ell = 1$. Further, if $T(t)$ is any function with a Taylor expansion, it follows as in the lemma that

$$(5.12)\quad \frac{1}{(\nu-1)!} \int_0^1 T(\mu h)\, \bar{G}^{[\ell]}(\mu h)\, \omega(x+\mu h)h\, d\mu = O(h^{2s+\beta-k+2}), \quad \ell=1.$$

Assume (5.11) and (5.12) are valid for $\ell-1 > 0$. Thus using (5.11), Lemma (5.1) implies that

$$\frac{1}{(\nu-1)!} \int_0^1 \bar{G}^{[\ell-1]}(\mu h)(h-h\mu)^{\nu-1}\, \omega(x+\mu h)h\, d\mu = O(h^{2s+\beta-k+2}) .$$

Then

$$\frac{1}{(\gamma-1)!}\int_0^1 T(\mu h)\,\bar{G}^{[\ell]}(\mu h)(h-h\mu)^{\gamma-1}\,\omega(x+\mu h)h\,d\mu$$

$$= \frac{1}{(\gamma-1)!}\int_0^1 T(\mu h)\left[\frac{1}{(m_\ell-1)!}\int_0^\mu \left\{\bar{G}^{[\ell-1]}(\mu_{\ell-1}h) - \sum_{\tau=s}^{2s-k-\gamma+1}\bar{G}^{[\ell-1](\tau)}(0)h^\tau(\mu_{\ell-1}^\tau\right.\right.$$

$$\left.\left. - \sum_{i_{\ell-1}=1}^{s} l_{i_{\ell-1}}(\mu_{\ell-1})\mu_{i_{\ell-1}}^\tau\right)\right\}(h\mu-h\mu_{\ell-1})^{m_\ell-1}\,\omega(x+\mu_{\ell-1}h)h\,d\mu_{\ell-1}\cdot$$

$$f(x+\mu h)_{m_\ell}\,\omega^{-1}(x+\mu h)\Big](h-\mu h)^{-1}\,\omega(x+\mu h)h\,d\mu$$

and as all integrals exist, the order of integration may be reversed to give

$$\frac{1}{(\gamma-1)!}\int_0^1 \frac{1}{(m_\ell-1)!}\left[\int_{\mu_{\ell-1}}^1 T(\mu h)(h\mu-h\mu_{\ell-1})^{m_\ell-1}f(x+\mu h)_{m_\ell}(h-\mu h)^{-1}\,h\,d\mu\right]\cdot$$

$$\left\{\bar{G}^{[\ell-1]}(\mu_{\ell-1}h) - \sum_{\tau=s}^{2s-k-\gamma+1}\bar{G}^{[\ell-1](\tau)}(0)h^\tau\left(\mu_{\ell-1}^\tau - \sum_{i_{\ell-1}=1}^{s} l_{i_{\ell-1}}(\mu_{\ell-1})\mu_{i_{\ell-1}}^\tau\right)\right\}\cdot$$

$$\omega(x+\mu_{\ell-1}h)h\,d\mu_{\ell-1}\,.$$

Now the integral in square brackets is a continuous function of $t = \mu_{\ell-1}h$, and has a Taylor series expansion in $[0,h]$. Therefore (5.12) implies that the total contribution corresponding to the first term in curlicue brackets is $O(h^{2s+\beta-k+2})$. For the remaining term, (5.7) may be applied, and each term of the summation becomes identically zero. Thus for $T(\mu h) = 1$, (5.11) is established for ℓ, and for $\gamma=1$, (5.12) is established for ℓ. Then (5.11) follows by an induction on ℓ, and the theorem is proved.

5. Numerical Example

Consider the problem

$$(5.10) \quad y' = \frac{t^{1/2}}{2}\left(7\sqrt{1 + t^{1/2}y} - 1\right)$$

$$y(0) = 0, \quad y'(0) = 0 .$$

Choosing the weight function $\omega(t) = t^{1/2}$, $g(0) = 3$ gives the solution

$$y(t) = 2t^{\frac{3}{2}} + t^{\frac{7}{2}} .$$

Now $g(t)$ and $\frac{\partial f}{\partial y} = \frac{7t}{4\sqrt{1+t^{1/2}y}}$ have Taylor series expansions about the origin; further the weight function satisfies (5.1) with $\beta = 1/2$, and thus the conditions of Theorem (5.1) are satisfied. Thus choosing the abscissae to be zeros of the polynomial orthogonal with respect to $t^{1/2}$ on $[0,1]$, methods of higher order are derived.

Using a variable weight function, the parameters must be recomputed at each step of the integration. As this computation is considerable, and even becomes unstable for later steps, it is desirable to use a unit weight function outside some neighbourhood of the singularity. In an attempt to determine when the unit weight function may replace $\omega(t) = t^{1/2}$ without a significant decrease in accuracy, accumulated errors were computed for both $\omega(t) = t^{1/2}$ and a unit weight function at each step using as initial values those previously calculated with $t^{1/2}$. For $s = 2,3,4$, and a steplength $h = .5$, these errors are tabulated in Table (5.1).

x	s	Zeros of $P_s(2\mu-1)$		Equally Spaced Abscissae	
		Wt. = $t^{1/2}$	Wt. = 1 at current step	Wt = $t^{1/2}$	Wt. =1 at current step
.5	2	-2.4 (-4)	-	-1.1 (-2)	-
	3	-5.5 (-12)	-	-3.6 (-12)	-
	4	-1.5 (-11)	-	5.5 (-12)	-
2.0	2	-1.3 (-3)	-1.6 (-3)	-1.8 (-1)	-2.2(-1)
	3	1.2 (-10)	1.0 (-6)	5.8 (-11)	2.5 (-4)
	4	4.7 (-10)	1.0 (-9)	1.7 (-10)	1.5 (-5)
4.0	2	-6.4 (-3)	-6.6 (-1)	-8.1 (-1)	-9.0 (-1)*
	3	2.0 (-10)	1.5 (-7)	-9.3 (-10)	8.5 (-5)
	4	2.0 (-10)	2.7 (-9)	4.2 (-9)	3.6 (-6)

TABLE (5.1): Errors in numerical solution of (5.10) with $h = .5$.

For $s = 2$, the expected difference in the order of the errors occurs. For this problem with $s = 3,4$, the choice of abscissae is immaterial as the solution is (effectively) a polynomial of degree three, and thus the errors correspond to rounding errors. Further the introduction of a unit weight function is not justified by the results at any step of integration for this problem for any value of s used.

---

* For this value the iteration scheme diverged.

CHAPTER VI

HYBRID METHODS

1. Introduction

Consistent single step methods are always stable, and s-stage implicit methods of order 2s exist. Although s-stage multistep methods (which are explicit) of order 2s are easily derived, Dahlquist [20] has shown that such methods of order greater than s+1 (or s+2 for s even) are unstable. To overcome this inadequacy, a modified version of these methods has been developed using additional function evaluations at each step.

Gragg and Stetter [22] developed the first methods of this type; using one arbitrary 'off-step' point, they derive several methods of maximum order. Gear [21] derives methods using a particular point, and indicates that either multi-step or these 'hybrid' methods are preferable to Runge-Kutta type methods for a well-behaved problem. Butcher [11] integrates a Hermite interpolation polynomial to derive a class of methods of high orders* using one arbitrary 'off-step' point. He derives a graph of the range of stability** for the choice of points in $[0,1]$. Later he uses the residue theorem from complex variable theory to derive methods using two or three 'off-step' points.

---

* The methods given by Butcher [11] have order less by 1 than the optimal methods given by Gragg and Stetter [22]. In the latter group of methods, the choice of the "off-step" point, μ_{s-1}, is severely restricted, and often $|\mu_{s-1}| > 1$.

** This graph is a plot of the magnitude of the largest eigenvalue other than unity of the matrix $A^{[1,0]}$ (defined by (2.6)) for these methods against the choice of μ_{s-1} for $s = 4(1)10$.

Using appropriate starting values (obtained, for example, using a single step method with a small steplength if necessary), these hybrid methods may be used to solve a system of first order differential equations. Here associated methods for a system of differential equations or arbitrary orders are developed. Indeed, two recursions analogous to (3.7) are obtained. As methods generated by the first recursion may be of lower order, and further, have a limited range of stability, they are in general inferior to those generated by the second.

2. Parameter Constraints

For a set of s abscissae, the first ℓ will denote those for which previously calculated values are used. Indeed, choose $\mu_i = \ell - i$, $i = 1(1)\ell$. Assuming that

$$\hat{y}_r^{(n_r-\nu)}(x_i) = y_r^{(n_r-\nu)}(x_i) + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = 1(1)\ell,$$

for(explicit) hybrid methods, new values are defined by

$$\hat{y}_r^{(n_r-\nu)}(x+\mu_i h) = \sum_{\sigma=0}^{\nu} \frac{(\mu_i h)^{\sigma}}{\sigma!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu,\sigma]} y_r^{(n_r-\nu+\sigma)}(x+\mu_j h),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1 ,$$

and

$$y_r^{(n_r)}(x+\mu_i h) = f_r(x_i ; \{\hat{y}_\rho^{(n_\rho-m)}\}) ,$$

$$r = 1(1)q, \quad i = \ell+1(1)s,$$

where we choose $\mu_{s+1} = 1$, so that the corresponding values are those required at $(x+h)$.

Similar to the approach in Chapter III, we proceed to establish conditions sufficient for a method to be of order p. Define

$$\psi_{ri}^{[\nu]} = \sum_{\tau=0}^{\nu} \frac{(\mu_i h)^{\tau}}{\tau!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu,\tau]} y_r^{(n_r-\nu+\tau)}(x_j) - y_r^{(n_r-\nu)}(x_i)$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1,$$

and a Taylor series expansion gives

$$\psi_{ri}^{[\nu]} = \sum_{\sigma=0}^{p} \frac{h^{\sigma}}{\sigma!} \left[\sum_{\tau=0}^{\nu} \binom{\sigma}{\tau} \mu_i^{\tau} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma} \right] y_r^{(n_r-\nu+\sigma)}(x) + O(h^{p+1}) . \tag{6.1}$$

Also define

$$\varepsilon_{ri}^{[\nu]} = y_r^{(n_r-\nu)}(x_i) - y_r^{(n_r-\nu)}(x_i) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1 ,$$

and

$$\varepsilon_{ri}^{[0]} = f_r(x_i ; \{\hat{y}_{\rho}^{(n_{\rho}-m)}(x_i)\}) - f_r(x_i ; \{y_{\rho}^{(n_{\rho}-m)}(x_i)\})$$

$$r = 1(1)q, \quad i = \ell+1(1)s.$$

Thus it follows that

$$(6.2)\quad \varepsilon_{ri}^{[\nu]} = \psi_{ri}^{[\nu]} + \sum_{\tau=0}^{\nu} \frac{(\mu_i h)^{\tau}}{\tau!} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu,\tau]}\, \varepsilon_{rj}^{[\nu-\tau]}$$

$$= \psi_{ri}^{[\nu]} + \sum_{\tau=0}^{\nu} \frac{(\mu_i h)^{\tau}}{\tau!} \sum_{j=\ell+1}^{i-1} \lambda_{rij}^{[\nu,\tau]}\, \varepsilon_{rj}^{[\nu-\tau]} + O(h^{p+1})$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1 \ ,$$

and

$$\varepsilon_{ri}^{[0]} = f_r(x_i\ ;\ \{y_\rho^{(n_\rho-m)}(x_i) + \psi_{\rho i}^{[m]} + \sum_{\tau=0}^{m} \frac{(\mu_i h)^{\tau}}{\tau!} \sum_{j=\ell+1}^{i-1} \lambda_{\rho ij}^{[m,\tau]}\, \varepsilon_{\rho j}^{[m-\tau]} + O(h^{p+1})\}\ )$$

$$- f_r(x_i\ ;\ \{y_\rho^{(n_\rho-m)}(x_i)\}\ )$$

$$r = 1(1)q, \quad i = \ell+1(1)s \ .$$

As before we assume the existence of first and second partial derivatives of the functions, and define

$$f_r(x_i)_{om} = \left[\frac{\partial f(t;\ \{z_\eta^{[\nu]}\})}{\partial z_\rho^{[m]}}\right]_{(x_i;\ \{y_\eta^{(n_\eta-\nu)}(x_i)\})} \ .$$

Then a Taylor series expansion truncated after second derivatives gives

$$(6.3)\quad \varepsilon_{ri}^{[0]} = \sum_{\rho=1}^{q} \sum_{m=1}^{n_\rho} f_r(x_i)_{\rho m} \left[\psi_{\rho i}^{[m]} + \sum_{\tau=0}^{m} \frac{(\mu_i h)^{\tau}}{\tau!} \sum_{j=\ell+1}^{i-1} \lambda_{\rho ij}^{[m,\tau]}\, \varepsilon_{\rho j}^{[m-\tau]}\right]$$

$$+ \frac{E_{ri}}{2} + O(h^{p+1})$$

$$r = 1(1)q, \quad i = \ell+1(1)s,$$

where E_{ri} is a sum of products of second derivatives with squares of terms given in square brackets.

Lemma (6.1): The parameters may be chosen so that

$$\psi_{ri}^{[\nu]} = O(h^{\ell}), \quad \varepsilon_{ri}^{[\nu]} = O(h^{\ell}),$$

for all appropriate indices. Further such a choice leads to

$$E_{ri} = O(h^{2\ell}), \qquad r = 1(1)q, \quad i = \ell+1(1)s.$$

Proof: As the abscissae μ_i, $i = 1(1)s$, are distinct, for any choice of $\lambda_{rij}^{[\nu,\tau]}$, $\tau > 0$, and $\lambda_{rij}^{[\nu,0]}$, $j > \ell$, the parameters $\lambda_{rij}^{[\nu,0]}$, $j \leq \ell$, may be chosen so that

$$\sum_{j=1}^{\ell} \lambda_{rij}^{[\nu,0]} \mu_j^{\sigma} = \mu_i^{\sigma} - \sum_{\tau=1}^{\nu} \binom{\sigma}{\tau} \mu_i^{\tau} \sum_{j=1}^{i-1} \lambda_{rij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \sum_{j=\ell+1}^{i-1} \lambda_{rij}^{[\nu,0]} \mu_j^{\sigma}$$

$$\sigma = 0(1)\ell-1, \quad r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1,$$

since the matrix of coefficients for each system of ℓ equations (given by a choice of r, ν, i) is non-singular. Thus

$$\psi_{ri}^{[\nu]} = O(h^{\ell}).$$

Now the assumption on the starting conditions and this result inply by a straightforward induction on i in (6.2) and (6.3) that

$$\varepsilon_{ri}^{[\nu]} = O(h^{\ell})$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad i = \ell+1(1)s+1,$$

and

$$\varepsilon_{ri}^{[0]} = O(h^{\ell})$$

$r = 1(1)q, \quad i = \ell+1(1)s.$

It follows that

$$E_{ri} = O(h^{2\ell})$$

for this choice of parameters.

Assume that

$$E_{ri} = O(h^{\bar{p}}), \qquad \bar{p} \geq 2\ell \ . \tag{6.4}$$

We wish to obtain methods of order $p \leq \bar{p}$, and thus by choosing a sufficient number of starting values (ℓ), methods of arbitrary order will be obtained.

Lemma (6.2): For a method to be of order p, $p \leq \bar{p}$, it is sufficient that

$$\sum_{\tau=0}^{\nu} \binom{\sigma}{\tau} \sum_{i=1}^{s} \alpha_{ri}^{[\nu,\tau]} \mu_i^{\sigma-\tau} = 1, \qquad \sigma = 0(1)p, \tag{6.5}$$

and

$$\sum_{i_0=1}^{s} \alpha_{r_0 i_0}^{[m_0,\tau_0]} \mu_{i_0}^{\sigma_1+\tau_1} \sum_{i_1=1}^{i_0-1} \lambda_{r_1 i_0 i_1}^{[m_1,\tau_1]} \cdots\cdots \mu_{i_{k-2}}^{\sigma_{k-1}+\tau_{k-1}}$$

$$\sum_{i_{k-1}=1}^{i_{k-2}-1} \lambda_{r_{k-1} i_{k-2} i_{k-1}}^{[m_{k-1},\tau_{k-1}]} \mu_{i_{k-1}}^{\sigma_k} \left[\sum_{\tau_k=0}^{m_k} \binom{\sigma}{\tau_k} \mu_{i_{k-1}}^{\tau_k} \sum_{i_k=1}^{i_{k-1}-1} \lambda_{r_k i_{k-1} i_k}^{[m_k,\tau_k]} \mu_{i_k}^{\sigma-\tau_k} - \mu_{i_{k-1}}^{\sigma} \right] = 0 \tag{6.5'}$$

for all positive integral k with

$$\tau_0 + \cdots + \tau_{k-1} + \sigma + \sigma_1 + \cdots + \sigma_k \le p, \quad \tau_j \ge 0, \quad \sigma_j \ge 0,$$

where

$$\sigma_j = 0, \quad m_j = m_{j-1} - \tau_{j-1}, \quad j = 1(1)k, \quad \text{unless } m_{j-1} = \tau_{j-1} .$$

Proof: Again we write $a_{ri}^{[\nu,\tau]}$ for $\lambda_{r\ s+1\ i}^{[\nu,\tau]}$. For a method to be of order p we must show that

$$\hat{y}_r^{(n_r-\nu)}(x+h) - y_r^{(n_r-\nu)}(x+h) = O(h^{p+1}),$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

Here these accumulated errors are given by (6.2):

$$\varepsilon_{r\ s+1}^{[\nu]} = \sum_{\sigma=0}^{p} \frac{h^\sigma}{\sigma!}\left[\sum_{\tau=0}^{\nu}\binom{\sigma}{\tau}\sum_{i=1}^{s} a_{ri}^{[\nu,\tau]}\,\mu_i^{\sigma-\tau} - 1\right] y_r^{(n_r-\nu+\sigma)}(x)$$

$$+ \sum_{\tau=0}^{\nu} \frac{h^\tau}{\tau!}\sum_{i=1}^{s} a_{ri}^{[\nu,\tau]}\,\varepsilon_{ri}^{[\nu-\tau]} + O(h^{p+1}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r .$$

Hence (6.5) implies we need only to show that

$$\sum_{i=1}^{s} a_{ri}^{[\nu,\tau]}\,\varepsilon_{ri}^{[\nu-\tau]} = O(h^{p+1-\tau}) ,$$

$$r = 1(1)q, \quad \nu = 1(1)n_r, \quad \tau = 0(1)\nu .$$

Now assume that the first order partial derivatives have Taylor series expansions. Then using (6.2) and (6.3) in one another recursively, we obtain an expansion of the errors $\varepsilon_{ri}^{[\nu-\tau]}$ in powers of h. For a method to be of order p,

it is sufficient that the coefficients of powers with index less than $p+1$ be zero. (Here, these coefficients are independent of h. Hence, for a method to be of order p, it is necessary that they be equal to zero. Indeed, using arguments similar to those following Lemma (3.2), it may be shown that (6.5) and (6.5') are necessary.)

In deriving terms recursively from (6.2), $\varepsilon_{rj}^{[\nu-\tau]}$ again requires (6.2), unless $\tau=\nu$, in which case (6.3) is required. As in Lemma (3.2), the coefficients of each power of h involve sums over the number of parameters $\{\lambda_{rij}^{[\nu,\tau]}\}$ occurring in each product, and the choices of τ_i, $i = 0(1)k$. It is sufficient for a method to be of order p, that each term of the sum be zero. A typical term of this sum is given by

$$\sum_{i_o=1}^{s} a_{r_o i_o}^{[m_o,\tau_o]} \left[f_{r_o}^{(\sigma_1)}(x)_{r_1 m_1} \frac{(\mu_{i_1} h)^{\sigma_1}}{\sigma_1!} \right] \frac{(\mu_{i_o} h)^{\tau_1}}{\tau_1!} \sum_{i_1=1}^{i_o-1} \lambda_{r_1 i_o i_1}^{[m_1,\tau_1]} \cdot$$

$$\left[f_{r_1}^{(\sigma_2)}(x)_{r_2 m_2} \frac{(\mu_{i_1} h)^{\sigma_2}}{\sigma_2!} \right] \frac{(\mu_{i_2} h)^{\tau_2}}{\tau_2!} \cdots\cdots \frac{(\mu_{i_{k-1}})^{\tau_{k-1}}}{\tau_{k-1}!} \sum_{i_{k-1}=1}^{i_{k-2}-1} \lambda_{r_{k-1} i_{k-2} i_{k-1}}^{[m_{k-1},\tau_{k-1}]} \cdot$$

$$\left[f_r^{(\sigma_k)}(x)_{r_k m_k} \frac{(\mu_{i_{k-1}} h)^{\sigma_k}}{\sigma_k!} \right] \frac{h^{\sigma}}{\sigma!} \left[\sum_{k=0}^{m_k} \binom{\sigma}{\tau_k} \mu_i^{\tau_k} \sum_{i_k=1}^{i_k-1} \lambda_{r_{k-1} i_k \mu_{i_k}}^{[m_k,\tau_k]} \sigma-\tau_k - \mu_{i_{k-1}}^{\sigma} \right]$$

where an expression in dotted brackets for k occurs only if $m_{k-1} = \tau_{k-1}$; otherwise $m_k = m_{k-1} - \tau_{k-1}$, and thus the statement of the lemma is valid.

3. Reduction to Methods for First Order Systems

Here, as in previous chapters, it is convenient to treat each equation in the same way. Thus parameters for a method for a system of equations are given, for example by $\lambda_{rij}^{[\nu,\tau]} = \lambda_{ij}^{[\nu,\tau]}$.

Theorem (6.1): If

$$(6.6)\qquad \sum_{\tau=0}^{l} \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[l,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma-\tau} = 0 ,$$

$$\sigma = 0(1)p, \qquad i = \ell+1(1)s+1 ,$$

and for $\nu > 1$, the following conditions are valid

$$(6.7)\quad \begin{cases} (1)\ \lambda_{ij}^{[\nu,\nu]} = 0, \quad j \neq \ell , \\ (2)\ \lambda_{ij}^{[\nu,\tau]}\mu_j + \mu_i \lambda_{ij}^{[\nu,\tau+1]} = \mu_i \lambda_{ij}^{[\nu-1,\tau]}, \quad j\neq\ell, \ \tau = 0(1)\nu-1 \\ (3)\ \lambda_{i\ell}^{[\nu,0]} = 1 - \sum_{j\neq\ell} \lambda_{ij}^{[\nu,0]} , \\ (4)\ \lambda_{i\ell}^{[\nu,\tau]} = \lambda_{i\ell}^{[\nu-1,\tau-1]} , \quad \tau = 1(1)\nu . \end{cases}$$

then for all ν,

$$(6.6')\qquad \sum_{\tau=0}^{\nu} \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma-\tau} = 0 .$$

$$\sigma = 0(1)p+\nu-1, \qquad i = \ell+1(1)s+1 .$$

Proof: For $\nu = 1$, (6.6) is identical to (6.6'). Then assume (6.6') is valid for all positive values less than some $\nu > 1$. Then we show this implies (6.6') for ν, and there are three cases to consider.

For $\sigma = 0$

$$\sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,0]} - 1 = 0$$

by (3). For $0 < \sigma \leq \nu$, certain well-known identities for the binomial coefficients give for all σ

$$\sum_{\tau=0}^{\sigma} \binom{\sigma}{\tau} \mu_i^{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma}$$

$$= \binom{\sigma-1}{0} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,0]} \mu_j^{\sigma} + \sum_{\tau=1}^{\sigma-1} \left[\binom{\sigma-1}{\tau-1} + \binom{\sigma-1}{\tau} \right] \mu_i^{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau}$$

$$+ \binom{\sigma-1}{\sigma-1} \mu_i^{\sigma} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\sigma]} - \mu_i^{\sigma}$$

$$= \sum_{\tau=0}^{\sigma-1} \binom{\sigma-1}{\tau} \sum_{j=1}^{i-1} \mu_i^{\tau} \left[\lambda_{ij}^{[\nu,\tau]} \mu_j + \mu_i \lambda_{ij}^{[\nu,\tau+1]} \right] \mu_j^{\sigma-1-\tau} - \mu_i^{\sigma} .$$

Using (2), and (4) for $\tau = \sigma - 1$, $j = \ell$, this is equal to

$$\sum_{\tau=0}^{\sigma-1} \binom{\sigma-1}{\tau} \sum_{j=1}^{i-1} \mu_i^{\tau} \left[\mu_i \lambda_{ij}^{[\nu-1,\tau]} \right] \mu_j^{\sigma-1-\tau} - \mu_i^{\sigma} = 0$$

since (6.6') is valid for $\nu - 1$. Similarly, for $\nu < \sigma \leq p+\nu-1$,

$$\sum_{\tau=0}^{\nu} \binom{\sigma}{\tau} \mu_i^{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma}$$

$$= \sum_{\tau=0}^{\nu-1} \binom{\sigma-1}{\tau} \sum_{j=1}^{i-1} \mu_i^{\tau} \left[\lambda_{ij}^{[\nu,\tau]} \mu_j + \mu_i \lambda_{ij}^{[\nu,\tau+1]} \right] \mu_j^{\sigma-1-\tau} - \mu_i^{\sigma} ,$$

using (1); as before this is equal to

$$\sum_{\tau=0}^{\nu-1} \binom{\sigma-1}{\tau} \sum_{j=1}^{i-1} \mu_i^{\tau} \left[\mu_i^{[\nu-1,\tau]} \lambda_{ij} \right] \mu_j^{\sigma-1-\tau} - \mu_i^{\sigma} = 0 \quad ,$$

$$\sigma - 1 \leq p + (\nu - 1) - 1 \quad ,$$

since (6.6') is valid for $\nu-1$. Now the result follows by induction on ν.

Now suppose that a method for a first order system of equations is chosen as in Lemma (6.1). Then the method is at least of order ℓ, and the recursion defined by the hypothesis of the theorem generates methods for systems of arbitrary orders in which lower-ordered derivatives are of <u>higher order</u>*. However, methods of order $p > \ell$ (and even $p = s+\ell-1$: Butcher [13]) exist, and thus methods for systems of arbitrary orders generated by this recursion may be of lower order than those from which they are generated. Indeed, if (6.5') with $m_j = 1$, $j = 0(1)k$, is valid for some value of p for which (6.6) is not valid, (6.5') may not be valid for this value of p for values of m_j greater than unity. This appears to be the case for the method using three 'off-step' points in the numerical example below.

Methods derived using this recursion have an additional limitation. Indeed, certain methods of high order (for example,

---

* This increase in order is a local phenomenon, for, errors in higher-ordered derivatives will eventually dominate the order of all components of the solution.

those given by Butcher [11, 13]) have a restricted range of stability. The stability of a method is determined by the parameters $\{\lambda_{ij}^{[\nu,0]}\}$. As (6.7) imply a change in these parameters for $\nu > 1$, methods for higher order equations may be unstable for all choices of 'off-step' points. Consider, for example, a method of order 5 (Butcher [11]) with $\ell=2$, $s=4$: $\lambda_{ij}^{[2,0]} = \lambda_{ij}^{[1,0]}$ and a method for a second order equation is stable whenever that for a first order equation is (at least for any choice of μ_3 in $[0,1]$). However, for $\nu=3$, methods are unstable for $\mu_3 < .8$, and for $\nu > 3$, all methods with $\mu_3 \in [0,1]$ are unstable.

The recursion given below generates methods which do not have these limitations.

Theorem (6.2): Suppose (6.5) and (6.5') are valid for p with

$$m_j = 1, \quad j = 0(1)k, \quad \text{and for } \nu > 1$$

$$(6.8)\quad \begin{cases} (5) & \lambda_{ij}^{[\nu,0]} = \lambda_{ij}^{[\nu-1,0]}, \\ (6) & \mu_i \lambda_{ij}^{[\nu,\tau]} = (\mu_i - \mu_j)\lambda_{ij}^{[\nu-1,\tau-1]}, \quad \tau = 1(1)\nu-1, \\ (7) & \mu_i \lambda_{ij}^{[\nu,\nu]} = \frac{\nu}{\nu-1}(\mu_i - \mu_j)\lambda_{ij}^{[\nu-1,\nu-1]}, \end{cases}$$

$$i = \ell+1(1)s+1, \qquad j = 1(1)i-1 .$$

Then (6.5) and (6.5') are valid for p for all values of m_j, $j = 0(1)k$.

Proof: Here (6.5) is given by (6.5') for $k = 0$, and thus we need only consider (6.5'). Assume (6.5') is valid for $m_i = 1$, $j = 0(1)k-1$, and $m_k = 1(1)\nu - 1$, and we prove (6.5') is valid for $m_k = \nu$. First, conditions (5) and (6) imply by an easy induction that

$$(8) \qquad \mu_i^{\tau} \lambda_{ij}^{[\nu,\tau]} = \mu_i^{\tau} \lambda_{ij}^{[\nu-1,\tau]}, \quad 0 \le \tau < \nu - 1.$$

Now we examine the expression in square brackets of (6.5'), and as in the previous theorem we must consider three cases. For $\sigma = 0$

$$\sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,0]} - 1 = \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu-1,0]} - 1$$

by (5), and as (6.5') is valid for $m_k = \nu-1$, it is valid for $m_k = \nu$. For $0 < \sigma < \nu$, the expression in square brackets may be written

$$\sum_{\tau=0}^{\sigma} \mu_i^{\tau} \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - \mu_i^{\sigma}$$

$$= \frac{1}{\nu-1}\left[\sum_{\tau=1}^{\sigma} \mu_i^{\tau} \tau \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} + \sum_{\tau=0}^{\sigma} \mu_i^{\tau} (\nu-1-\tau) \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau]} \mu_j^{\sigma-\tau} - (\sigma + \nu-1-\sigma)\mu_i^{\sigma} \right].$$

Using (8) in the second sum, this is equal to

$$\frac{1}{\nu-1}\left[\sum_{\tau=0}^{\sigma-1} \mu_i^{\tau+1} (\tau+1) \binom{\sigma}{\tau+1} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu,\tau+1]} \mu_j^{\sigma-\tau-1} - \sigma\mu_i^{\sigma} + \sum_{\tau=0}^{\sigma} \mu_i^{\tau} (\nu-1-\tau) \binom{\sigma}{\tau} \sum_{j=1}^{i-1} \lambda_{ij}^{[\nu-1,\tau]} \mu_j^{\sigma-\tau} - (\nu-1-\sigma)\mu_i^{\sigma} \right].$$

since the only term with $\tau \geq \nu-1$ has $\tau = \sigma = \nu-1$, and the corresponding contribution is zero. Using (6) in the first sum, further rearrangement gives

$$\frac{1}{\nu-1}\left[\sigma\left[\sum_{\tau=0}^{\sigma-1}\mu_i^{\tau}\binom{\sigma-1}{\tau}\sum_{j=1}^{i-1}\lambda_{ij}^{[\nu-1,\tau]}(\mu_i-\mu_j)\mu_j^{\sigma-\tau-1} - \mu_i^{\sigma}\right]\right.$$

$$\left.+\left[\sum_{\tau=0}^{\sigma}\mu_i^{\tau}(\nu-1-\tau)\binom{\sigma}{\tau}\sum_{j=1}^{i-1}\lambda_{ij}^{[\nu-1,\tau]}\mu_j^{\sigma-\tau} - (\nu-1-\sigma)\mu_i^{\sigma}\right]\right]$$

$$= \frac{\sigma\mu_i}{\nu-1}\left[\sum_{\tau=0}^{\sigma-1}\mu_i^{\tau}\binom{\sigma-1}{\tau}\sum_{j=1}^{i-1}\lambda_{ij}^{[\nu-1,\tau]}\mu_j^{\sigma-1-\tau} - \mu_i^{\sigma-1}\right]$$

$$+ \frac{\nu-1-\sigma}{\nu-1}\left[\sum_{\tau=0}^{\sigma}\mu_i^{\tau}\binom{\sigma}{\tau}\sum_{j=1}^{i-1}\lambda_{ij}^{[\nu-1,\tau]}\mu_j^{\sigma-\tau} - \mu_i^{\sigma}\right] .$$

As (6.5') is valid for $m_k = \nu-1$, the latter expression implies that it is valid for $m_k = \nu$. For $\sigma \geq \nu$, the proof is similar to that above, using (6) and (7) in place of (6). Then it follows that (6.5') is valid for all values of m_k by induction.

The result of the theorem now follows by induction on m_j, $j = 0(1)k-1$. For suppose (6.5') is valid for $m_j > 0$, $j \neq i$, and $m_i = \nu_i - 1$. Then one of (5), (6) or (7) may be used to establish (6.5') for $m_i = \nu_i$ by reducing it to one or more similar expressions for $m_i = \nu_i - 1$.

Thus (6.8) generates a method of order p for systems of arbitrary orders from one of order p for a first order

system. (However, there is not the increase in (local) order for lower-ordered derivatives provided by (6.7).) Further (5) implies that a method for higher-ordered systems is stable whenever that for first order systems from which it was generated is.

Theorem (6.3): To obtain an s-stage method of order p, choose ℓ so that $p \leq 4\ell$. If

$$\psi_{ri}^{[\nu]} = O(h^{2\ell}), \quad i = \ell+1(1)s+1,$$

then (6.5) and (6.5') are necessary and sufficient for a method to be of order p, $p > 2\ell$.

Proof: Sufficiency is implied by Lemma (6.2).

Butcher [11, 13] shows that the parameters may be chosen so that

$$\psi_{r1}^{[\nu]} = O(h^{2\ell}), \quad i = \ell+1(1)s+1 \ .$$

Indeed, with ℓ solution values and ℓ derivatives, a Hermite polynomial may be used to approximate a new value, and this is exact for all polynomials of degree less than 2ℓ. The result follows, for example, by integrating the Hermite polynomial to obtain the parameters. As before $E_{ri} = O(h^{4\ell})$, and for $p \leq 4\ell$, it is necessary and sufficient that the accumulated errors be of order p. From the discussion for Lemma (6.2), the necessity and sufficienty of (6.5) and (6.5') follows.

Butcher [11, 13] gives methods of orders $p = \ell+s-1 > 2\ell$.

(Further, he proved that no method exists with $\ell = 1$, $s-\ell-1 = 3$, in which case $\ell+s-1 > 4\ell$.) It follows from the theory that these methods satisfy (6.5) and (6.5') for $p = \ell+s-1$, and they may be used with (6.8) to generate methods of order p for systems of arbitrary orders.

4. Numerical Example

Here we solve a second order system

$$(6.9) \quad y_1' = y_2y_3, \quad y_2' = y_1y_3', \quad y_3'' = \sqrt{-y_3y_3'}$$

which has the solution (for appropriate initial values)

$$y_1 = \sin(\exp(-x)), \quad y_2 = \cos(\exp(-x)), \quad y_3 = \exp(-x),$$

and the first order system derived therefrom

$$(6.10) \quad y_1' = y_2y_3, \quad y_2' = y_1y_4, \quad y_3' = y_4, \quad y_4' = \sqrt{-y_3y_4} \ .$$

Recursions (6.7) and (6.8) are used to derive methods for the solution of (6.9) from several methods given by Butcher [11,13] for first order systems. A comparison is made of the time required for, and errors given by, these methods and explicit Runge-Kutta methods given by Lawson* [29,30].

In Table (6.1), the methods used are listed, and the (minimum) orders are tabulated.

---

* The method of order 5 quoted [29, p. 597] should have $\mu_3\lambda_{31} = \frac{3}{16}$, $\mu_3\lambda_{32} = \frac{1}{16}$, not $\mu_3\lambda_{31} = \mu_3\lambda_{32} = \frac{1}{8}$.

Symbol	Source	ℓ	$s+\ell-1$	Order		
				Problem (6.9)		
				Recursion (6.7)	Recursion (6.8)	Problem (6.10)
L5	Lawson [29]	1	4	2	5	5
L6	Lawson [30]	1	5	2	6	6
B5	Butcher [11]	2	1	5	5	5
B7	Butcher [11]	3	1	7	7	7
$\hat{B}7$	Butcher [13]	2	3	5	7	7
B9	Butcher [11]	4	1	9	9	9
B11	Butcher [11]	5	1	11	11	11

TABLE (6.1): The orders of certain methods for problems solved.

Method	Problem and Recursion		
	(6.9)		(6.10)
	(6.7)	(6.8)	
L5	-	242.6308	224.2665
L6	-	308.0271	283.5278
B5	117.3570	117.3620	110.1482
B7	143.0687	143.0608	133.9037
$\hat{B}7$	225.6817	225.6989	208.9102
B9	168.7994	168.7826	157.6789
B11	194.5331	194.5076	181.4527

TABLE (6.3): Time in seconds for 250 steps.

Method	$\varepsilon_1^{[1]}$			$\varepsilon_2^{[1]}$		
	(6.9)		(6.10)	(6.9)		(6.10)
	(6.7)	(6.8)		(6.7)	(6.8)	
L5	-	-2.2(-11)	2.1(-12)	-	-3.2(-11)	-2.9(-11)
L6	-	1.0(-13)	7.9(-13)	-	-3.0(-13)	9.4(-12)
B5	1.8(-10)	1.3(-10)	7.0(-11)	-9.0(-10)	-1.0(-9)	-8.3(-10)
B7	1.9(-12)	1.9(-12)	1.8(-12)	-7.7(-13)	-8.3(-13)	-7.6(-13)
$\hat{B}7$	-6.4(-10)	1.4(-12)	1.1(-12)	-3.0(-13)	2.8(-13)	2.7(-13)
B9	1.1(-15)	1.1(-15)	1.1(-15)	2.2(-15)	2.2(-15)	2.2(-15)
B11	-4.9(-18)	-4.9(-18)	-4.9(-18)	7.1(-18)	7.1(-18)	7.1(-18)
	$\varepsilon_3^{[2]}$			$\varepsilon_3^{[1]}$		
L5	-	-2.0(-12)	-6.8(-12)	-	1.0(-12)	6.8(-12)
L6	-	-1.5(-14)	3.1(-14)	-	5.3(-16)	-3.1(-14)
B5	-2.7(-11)	6.0(-11)	-1.0(-10)	1.1(-10)	-1.2(-10)	1.0(-10)
B7	-5.6(-15)	5.2(-14)	-3.2(-14)	3.9(-14)	-5.3(-14)	3.2(-14)
$\hat{B}7$	-2.5(-10)	-1.9(-13)	6.8(-14)	-4.3(-10)	-1.2(-14)	-6.8(-14)
B9	-6.0(-19)	1.3(-17)	-3.4(-18)	3.7(-18)	-5.3(-18)	3.4(-18)
B11	2.0(-21)	3.1(-21)	-8.0(-22)	6.4(-22)	-1.5(-21)	4.0(-22)

TABLE (6.2): Errors at x = 1 in the solution of problems (6.9) and (6.10) on [0, 1] with h = .03125 .

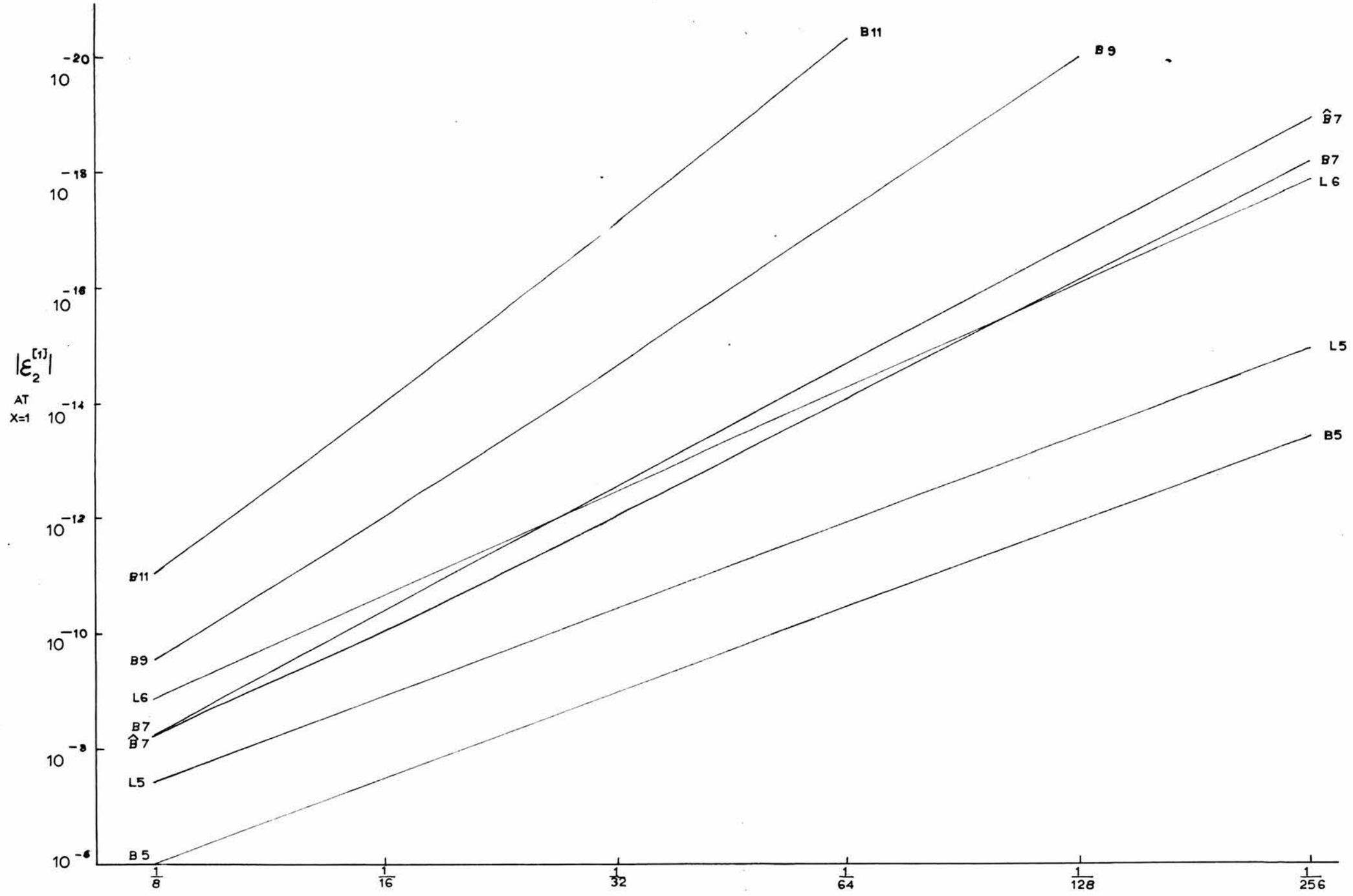

FIG. (6·1) : ERROR – STEPLENGTH GRAPH FOR PROBLEM (6·9) USING RECURSION (6·8)

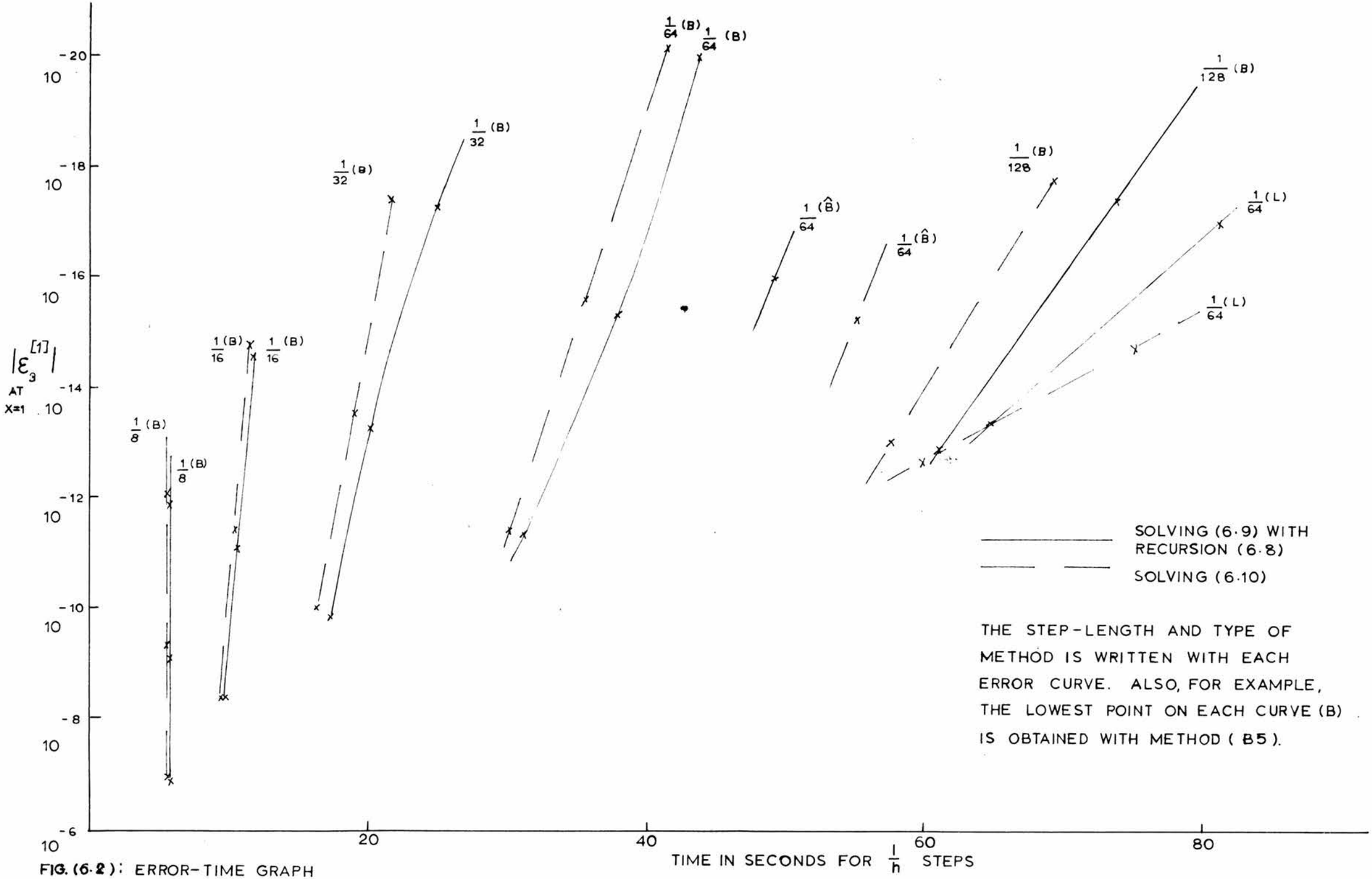

FIG. (6·2): ERROR-TIME GRAPH

The problems are solved using exact starting values for each method, and for $h = \frac{1}{32}$, the errors at $x = 1$ are tabulated in Table (6.2). The third table gives the time required for 250 steps for each method. A further comparison is given in two graphs in which the error at $x = 1$ is plotted against steplength and time respectively for methods used.

The results of Table (6.2), in general, support the theory of the foregoing sections. As the accuracies for the solution of (6.9) using (6.8), and (6.10) are comparable, Table (6.3) indicates for this problem methods for (6.10) are slightly more economical. (As methods for (6.9) generated by (6.7) are inferior with respect to accuracy and stability, it appears that they are of no practical use.) However, for problems such as those examined by de Vogeleare [41], in which lower-ordered derivatives are not required in the function evaluation, methods generated by (6.8) may be more economical.

Figure (6.1) provides a comparison of the error versus the steplength for the methods examined for problem (6.9). Figure (6.2) indicates (for this problem at least) that the most economical of the methods examined is a hybrid method of high order using one 'off-step' point with a large steplength.

CHAPTER VII

QUADRATURES FOR IMPLICIT DIFFERENTIAL EQUATIONS

1. Introduction

Methods for the numerical solution of a system of differential equations are usually based on the assumption that the equations are respectively explicit in each of the highest-ordered derivatives. If this is not the case, the equations may be differentiated with respect to the independent variable provided that the new derivatives are sufficiently well-behaved (Collatz [15, p. 97]). This yields another implicit system which is linear in the new (highest-ordered) derivatives, and may be solved explicitly for these derivatives provided that the matrix of coefficients is non-singular.

It seems reasonable to determine if there exist more direct methods for solving such problems. Indeed such methods are required if the technique outlined above is not valid. As the behaviour of higher-ordered derivatives is not known or may not easily be determined in general, it would be prudent to use direct methods if they were available. Thus this chapter is principally concerned with the development of a computationally efficient numerical method for solving implicit differential equations directly. As a result of this investigation a useful extension of an existence theorem for implicit differential equations is obtained.

As seen in the preceding chapters, the essence of a numerical method is the replacement of the differential system by an algebraic

system. The solution of this system provides numerical approximations to the highest-ordered derivatives for several values of the independent variable. The algebraic system derived from an implicit differential equation will, in general, be non-linear, and may be derived using any suitable method for an initial value problem for explicit differential equations. Only single step methods are considered, and, for example, an explicit Runge-Kutta method could be used. However a smaller algebraic system arises by using an implicit method (of the same order of accuracy) of the type examined in Chapter IV, and as the system must usually be solved by an iterative technique, it appears more economical to use a method of this class. Indeed, a suitable method will be chosen from the more general class of methods examined in Chapter V. There remains then the problem of developing a reasonably efficient iterative scheme for the particular solution of the algebraic system required.

Thus we shall consider a differential system of the form (1.1). Before describing the methods in detail, certain existence theorems are investigated.

2. Existence Theorems

The introduction of a weight function leads to some interesting questions of existence, even in the case of explicit differential equations. For example, if $f(t,y)$ is a continuous function in an open neighbourhood of the origin, does there exist a unique solution to the problem

$$y' = t^{-\alpha} f(t, y), \qquad 0 < \alpha < 1$$

$$y(0) = 0, \quad f(0,0) = f_0 \neq 0,$$

in some interval $[0,a]$, $a > 0$? Although we shall not attempt to answer this question, we note (with some interest) that, with the discontinuous weight function $\omega(t) = t^{-\alpha}$, a numerical solution for this problem can be determined using a quadrature method of Chapter V.

An analogous question arises for implicit differential equations. Indeed, it is shown below under certain conditions that the implicit differential problem is equivalent to an explicit one, and has a unique solution if and only if one exists for the explicit problem. As the methods described provide numerical solutions to an implicit differential problem with a discontinuity similar to that of the explicit problem above, this analogy seems reasonable.

First, using continuous weight functions, the existence of solutions for certain implicit differential equations may be proved, and we begin by modifying the statement of the implicit function theorem. Consider a partially open (non-trivial) region D of real Euclidean space R_{m+n+1}:

$(t; \underline{u}, \underline{v})$ lies in D if $0 \leq t - x < a$,

$$|u_i - u_{io}| < a, \quad i = 1(1)m,$$

$$|v_i - v_{io}| < a, \quad i = 1(1)n,$$

for some point $(x; \underline{u}_o, \underline{v}_o)$ of the space.

Then with appropriate adjustments to their proofs, Theorems 1 and 3 given by Murray and Miller [5, pp. 22, 30] may be restated as

Theorem (7.1)

H1: Let

$$G_i(t; \underline{u}, \underline{v}), \qquad i = 1(1)n,$$

be a set of real-valued functions of $(m+n+1)$ real variables, defined and continuous on D of R_{m+n+1}.

H2: Let these functions have continuous first derivatives relative to the variables v_j, $j = 1(1)n$, at every point of D.

H3: Let $(x; \underline{u}_o, \underline{v}_o)$ be a point of D such that

$$G_i(x; \underline{u}_o, \underline{v}_o) = 0, \qquad i = 1(1)n,$$

and the Jacobian

$$J = \frac{\partial(G_1, \ldots, G_n)}{\partial(v_1, \ldots, v_n)}$$

is not zero at this point.

H4: For some fixed value of j assume $\frac{\partial G_i}{\partial u_j}$ exist and are continuous for $i = 1(1)n$.

Then under hypotheses H1, H2, and H3 there exists a positive number b a and n continuous functions $\phi_i(t; \underline{u})$ such that

$$v_{io} = \phi_i(x; \underline{u}_o), \qquad i = 1(1)n,$$

and

$$G_i(t; \underline{u}, \underline{\phi}) = 0, \qquad i = 1(1)n,$$

for a vector $\underline{u}$ such that

$$|u_k - u_{ko}| \leq b, \qquad k = 1(1)m, \qquad 0 \leq t-x \leq b .$$

In addition, under hypothesis H4, the partial derivatives $\frac{\partial \phi_i}{\partial u_j}$ exist and are continuous for these values of $\underline{u}$.

This theorem leads to the following extension of the existence theorem for the solution of implicit differential equations given by Murray and Miller [5, p. 32]. The adjustments to their proof follow the statement of the theorem. Consider a partially open region D of real Euclidean space R_{N+q+1}:

$(t; \underline{y}, \underline{g})$ lies in D if $0 \leq t < a$,

$$\left| y_r^{(n_r-m)} - y_{ro}^{(n_r-m)} \right| < a, \quad r = 1(1)q, \quad m = 1(1)n_r,$$

$$\left| g_r - g_{ro} \right| < a, \quad r = 1(1)q.$$

Theorem (7.2): Define $\omega_r(t)$, $r = 1(1)q$, a set of weight functions which are continuous in the half-open interval $[x, a)$, and a vector $\underline{g}_o$ such that

$$z_{ro} = g_{ro}\, \omega_r(0), \quad r = 1(1)q.$$

Consider the functions

$$F_r^{\star}(t; \underline{y}, \underline{g}) = F_r(t; \underline{y}, \{\omega_\rho(t) g_\rho\}), \quad r = 1(1)q,$$

in the region D. Suppose that

H1: $F_r^{\star}(t; \underline{y}, \underline{g})$ are continuous in D for $r = 1(1)q$.

H2: $\frac{\partial F_r^{\star}}{\partial g_\rho}(t; \underline{y}, \underline{g})$ exist and are continuous on D for $r, \rho = 1(1)q$.

H3: $F_r^{\star}(x; \underline{y}_o, \underline{g}_o) = 0$, and the Jacobian

$$J = \frac{\partial(F_1^{\star}, \ldots, F_q^{\star})}{\partial(g_1, \ldots, g_q)}$$

is not zero at the point $(x; \underline{y}_o, \underline{g}_o)$.

H4: $\dfrac{\partial F_r^*}{\partial y_\rho^{(n_\rho - m)}}(t; \underline{y}, \underline{g})$ exist and are continuous on D for

$$r, \rho = 1(1)q, \quad m = 1(1)n_\rho .$$

Then there exists a number $b > x$ such that (1) has a solution $\underline{y}(t)$ which has continuous first derivatives in $[x, b]$, and

$$\underline{y}(x) = \underline{Y}_o, \qquad \underline{z}(x) = \underline{z}_o ;$$

furthermore the solution is unique for each vector $\underline{g}_o$ satisfying the hypotheses.

For each choice of $\underline{g}_o$ Theorem (7.1) implies the existence of q continuous functions

$$g_r = \phi_r(t; \underline{y}), \qquad r = 1(1)q,$$

such that

$$F_r^*(t; \underline{y}, \phi) = 0. \qquad r = 1(1)q.$$

As the weight functions are continuous so also are the functions

$$y_r^{(n_r)} = \omega_r(t)\, g_r = \omega_r(t)\, \phi_r(t; \underline{y}), \quad r = 1(1)q,$$

and existence follows. Uniqueness follows from a Lipschitz condition on the functions $\phi_r(t; \underline{y})$ (as that shown to exist by Murray and Miller [5, p. 45]), and the boundedness of $\omega_r(t)$ in some subinterval of $[x, a)$ containing x.

For example, consider the problem

$$\frac{(y')^2}{t} \cos \frac{y'}{t^{\frac{1}{2}}} - y + 1 = 0,$$

$$y(0) = 1, \quad y'(0) = 0 .$$

It does not appear that we may apply the existence theorem given by Murray and Miller. However, by choosing $\omega(t) = t^{\frac{1}{2}}$, and $g_o = (k+\frac{1}{2})\pi$, $k = 0, \pm 1, \pm 2, \ldots$ so that

$$y'(0) = g_o\,\omega(0) = 0,$$

$$F(t; \underline{y}, g) = g_o^2 \cos g_o - y_o + 1 = 0,$$
$$(0; \underline{y}_o, g_o)$$

$$\frac{\partial F}{\partial g}(t; \underline{y}, g) = g_o(2\cos g_o - g_o \sin g_o) \neq 0,$$
$$(0; \underline{y}_o, g_o)$$

Theorem (7.2) guarantees a unique solution for each value of g_o.

This approach is also useful in considering an implicit differential system for which the hypotheses of Theorem (7.2) may be satisfied only by weight functions which are discontinuous at the origin. The proof again leads to the existence of continuous functions $\phi_r(t; \underline{y})$ as above, but as the weight functions are discontinuous, the question of existence of a solution to the differential problem

$$y_r^{(n_r)} = \omega_r(t)\,\phi_r(t; \underline{y}), \qquad r = 1(1)q,$$

is a generalization of that asked at the beginning of this section.

When a unique solution does exist for a problem associated with either continuous or discontinuous weight functions in this way, it will be shown using Theorem (7.1) that the numerical methods proposed are convergent under suitable conditions. Henceforth we assume that there exists a unique solution $\underline{y}(t)$ of (1.1),

continuous on some interval $I = [x, a]$ of the real variable t, satisfying the initial conditions at $t = x$ (and, if necessary, associated with some particular vector g_o as given in Theorem (7.2)). For some positive integer p $(p \geq \max n_r)$, it is assumed that there exists a set of weight functions $\omega_r(t)$, $r = 1(1)q$, such that the derivatives

$$(7.1) \qquad \frac{d^{p-n_r}}{dt^{p-n_r}} \{\omega_r^{-1}(t) y_r^{(n_r)}(t)\}, \qquad r = 1(1)q,$$

are continuous in I.

A numerical solution of (1.1) is determined when approximations $\hat{y}(x+h)$ to $y(x+h)$ are defined, where h, the steplength, is some interval length of the real variable t, chosen so that $x+h$ lies in I.

3. Numerical Methods and Convergence

Here quadrature methods are applied to obtain a numerical solution of (1.1). Indeed, for a set of distinct abscissae, let the parameters be derived using the quadrature formulae of Chapter V. Then approximations to $g_r(x_i)$,

$$g_{ri} = g_r(x_i) + \varepsilon_{ri} \qquad r = 1(1)q, \quad i = 1(1)s,$$

are required, and these are defined implicitly by

$$(7.2) \qquad F_r(x_i; \{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij} g_{\rho j}\}, \{\omega_\rho(x_i) g_{\rho i}\}) = 0,$$

$$i = 1(1)s, \quad r = 1(1)q.$$

In general equations (7.2) are non-linear, and the approximations $\{g_{ri} \quad i = 1(1)s, \quad r = 1(1)q\}$ have to be obtained iteratively; a suitable iteration scheme is described in the next section. Provided that the errors $\{\varepsilon_{ri}\}$ are small, these approximations provide adequate approximations $\hat{\underline{y}}(x+h)$ to $\underline{y}(x+h)$ by using (5.2) as described in Chapter V. It will now be shown that h may be chosen sufficiently small that the errors $\{\varepsilon_{ri}\}$ may be neglected, and further that the method converges.

Theorem (7.3): Suppose there exists a continuous solution of (1.1) which is unique to a choice of g_o. Let the hypotheses of Theorem (7.2) be satisfied for a set of weight functions which provide the continuous derivatives given by (7.1). Then there exist continuous functions $g_{ri}(h)$, $i = 1(1)s$, $r = 1(1)q$, which satisfy (7.2) and such that

$$g_{ri}(h) = g_r(x_i) + O(h^{s+1})$$

for sufficiently small h.

Proof: (i) By Theorem (7.2) there exist continuous functions

$$g_r = \phi_r(t; \underline{y}), \quad r = 1(1)q,$$

and as there exists a continuous solution to the differential problem, these functions may be considered as continuous functions of t on $I' = [x, b]$, b a, and further they satisfy

$$F_r^*(t; \underline{y}(t), \underline{g}(t)) = 0, \quad r = 1(1)q.$$

As in the proof of the implicit function theorem given by Murray and Miller [5, p. 23], there exists a (convex) neighbourhood N

of $(x;\ y_o,\ g_o)$ in D such that the Jacobian

$$J = \frac{\partial(F_1^*, \ldots, F_q^*)}{\partial(g_1, \ldots, g_q)}$$

is not zero in this neighbourhood.

(ii) There exists a continuous solution $\{g_{ri}(h),\ i = 1(1)s,\ r = 1(1)q\}$ of (7.2). For, consider the equations

$$\hat{F}_{ri}(h;\ g_1, \ldots, g_s) \equiv F_r^*\left(x_i;\ \{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}\, g_{\rho j}\},\ \{g_{\rho i}\}\right) = 0,$$

$$r = 1(1)q, \qquad i = 1(1)s,$$

and a region E of R_{qs+1} :

$(h;\ g_1, \ldots, g_s)$ lies in E if $0 \le h < a$,

$$|g_{ri} - g_{ro}| < a, \quad r = 1(1)q,\ i = 1(1)s.$$

H1: By H1 of Theorem (7.2), $\hat{F}_{ri}(h;\ g_1, \ldots, g_s)$, $r = 1(1)q$, $i = 1(1)s$, are continuous in some neighbourhood M_1 of $(0;\ g_o, \ldots, g_o)$ contained in E .

H2: By H2 and H4 of Theorem (7.2) there exists a neighbourhood M_2 of $(0;\ g_o, \ldots, g_o)$ contained in E in which $\frac{\partial \hat{F}_{ri}}{\partial g_{\rho j}}(h; g_1, \ldots, g_s)$ exist and are continuous.

H3: By H3 of Theorem (7.2),

$$\hat{F}_{ri}(0;\ g_o, \ldots, g_o) = F_r^*\left(x;\ \{y_{\rho o}^{(n_\rho - m)}\},\ \{g_{\rho o}\}\right) = 0 .$$

Further, the Jacobian matrix

$$\{J\} = \left\{ \frac{\partial(\hat{F}_{11},\ldots,\hat{F}_{q1};\ldots;\hat{F}_{1s},\ldots,\hat{F}_{qs})}{\partial(g_{11},\ldots,g_{q1};\ldots;g_{1s},\ldots,g_{qs})} \right\}$$

at the point $(0; g_o,\ldots, g_o)$ has non-zero elements only in the $q \times q$ sub-matrices on the diagonal

$$\{J_{ii}\} = \left\{ \frac{\partial(\hat{F}_{1i},\ldots,\hat{F}_{qi})}{\partial(g_{1i},\ldots,g_{qi})} \right\}, \qquad i = 1(1)s,$$

each of which has a non-zero determinant at $(0; g_o,\ldots, g_o)$ since it is identical with the Jacobian of H3 in Theorem (7.2). We remark that off-diagonal sub-matrices occur as powers of h, and thus vanish at the point $(0; g_o,\ldots, g_o)$.

Thus by Theorem (7.1) there exists a positive number h_o and qs continuous functions

$$g_{ri}(h), \qquad r = 1(1)q, \quad i = 1(1)s,$$

for $0 \le h < h_o$ such that

$$g_{ri}(0) = g_{ro}, \qquad r = 1(1)q, \quad i = 1(1)s,$$

and

$$\hat{F}_{ri}(h; g_1(h),\ldots, g_s(h)) = 0, \quad r = 1(1)q, \quad i = 1(1)s.$$

Further these functions are unique in some region $E' \subseteq E$.

(iii) Since $g_{ri}(h)$, $r = 1(1)q$, $i = 1(1)s$, are continuous, there exists a positive number $h_1 < h_o$ such that for $0 \le h \le h_1$, the point

$$(x_i;\ \underline{Y}_i,\ \underline{g}_i) = (x_i;\ \{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}^{[m]}\, g_{\rho j}\},\ \{g_{\rho i}\})$$

lies in the neighbourhood N of (i). Then let h be chosen so that $0 \leq h \leq h_1$, and further so that $(x_i;\ \underline{Y}(x_i),\ \underline{g}(x_i))$ lies in N. As N is convex, any point $(x_i;\ \underline{Y}_i^*,\ \underline{g}_i^*)$ whose components lie between those of $(x_i;\ \underline{Y}_i,\ \underline{g}_i)$ and $(x_i;\ \underline{Y}(x_i), \underline{g}(x_i))$ is a point of N. Now (7.2) may be expanded in a Taylor series. This expansion is equivalent to

$$(7.3)\quad 0 = F_r^*(x_i;\ \{y_\rho^{(n_\rho - m)}(x_i)\},\ \{g_\rho(x_i)\})$$

$$+ \sum_{\rho=1}^{q}\sum_{m=1}^{n_\rho}\left(-R_{\rho i}^{[m]} + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}^{[m]}\varepsilon_{\rho j}\right)\frac{\partial F_r^*}{\partial y_\rho^{(n_\rho - m)}}\Bigg|_{(x_i;\underline{Y}_1^*,\underline{g}_i^*)}$$

$$+ \sum_{\rho=1}^{q}\varepsilon_{\rho i}\left.\frac{\partial F_r^*}{\partial g_\rho}\right|_{(x_i;\ \underline{Y}_i^*,\ \underline{g}_i^*)},\qquad r = 1(1)q,\quad i = 1(1)s,$$

where $(x_i;\ \underline{Y}_i^*,\ \underline{g}_i^*)$ lies in N.

By (i) the first term is identically zero. Further the Jacobian matrix

$$\{J_i^*\} = \left\{\frac{\partial(F_1^*, \ldots, F_q^*)}{\partial(g_1, \ldots, g_q)}\right\}_{(x_i;\ \underline{Y}_i^*,\ \underline{g}_i^*)}$$

has a non-zero determinant. (7.3) may be written in matrix notation as

$$(J^* + hZ)\ \underline{\varepsilon} = \underline{\psi}$$

where

$$J^* = \begin{pmatrix} \{J_1^*\} & . & . & . & 0 \\ . & \{J_2^*\} & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \{J_q^*\} \end{pmatrix}$$

has a non-zero determinant, Z is a continuous function of h, and thus h may be chosen smaller if necessary so that $(J^* + hZ)$ is non-singular. Further

$$\underline{\varepsilon}^T = (\varepsilon_{11},\ldots, \varepsilon_{q1}\ ;\ldots;\ \varepsilon_{1s},\ldots, \varepsilon_{qs})$$

and the parameters may be chosen so that

$$\underline{\Psi} = O(h^{s+1}) \quad .$$

Thus for small enough h,

$$\underline{\varepsilon} = (J^* + hZ)^{-1}\underline{\Psi} = O(h^{s+1})$$

and the theorem is proved.

To prove convergence, a little additional work is required. The expression $T_\rho^{[m]}(\mu_i h)$ of (7.2) must be adjusted to include the accumulated errors; a smiliar adjustment to $R_{\rho i}^{[m]}$ of (7.3) will account for this error in the proof of Theorem (7.3), and the result of the theorem must be correspondingly restated. Convergence of the method then follows by a proof similar to that given by Cooper [16].

As the weight functions need not be continuous for the proof

of the theorem, good numerical solutions may be obtained for certain problems in which the highest-ordered derivatives are discontinuous.

4. Iteration Scheme

As (7.2) is a non-linear system of algebraic equations, it will, in general, have more than one solution in R_{sq}. The unique solution in E' given by Theorem (7.3) is required, and this may not be determined by a general iterative technique for (7.2). We now restrict attention to a system of the form (1.1) in which $y_r^{(n_r)}(t)$ occurs only in the r-th equation, and in this case the Jacobian matrix of Theorem (7.3) is diagonal. Successive iterates will be denoted with superfixes, and for some fixed steplength h, a generalized Newton method gives

$$(7.4)\quad g_{ri}^{(k+1)} = g_{ri}^{(k)} - C_{ri}^{(k)}\, F_r^{*}\Big(x_i;\ T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}\, g_{\rho j}^{(k)}\Big\},\ g_{ri}^{(k)}\Big)$$

$$r = 1(1)q, \quad i = 1(1)s,$$

where

$$C_{ri}^{(k)} = \frac{1}{\left.\dfrac{\partial F_r}{\partial g_r}\right|_{\left(x_i;\ \left\{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}\, g_{\rho j}^{(k)}\right\},\ g_{ri}^{(k)}\right)}}.$$

If $C_{ri}^{(k)}$ is small (< 1), it may be accurately approximated

by $C_{ri}^{(o)}$ throughout the iteration. This is not true for larger values. Indeed if $C_{ri}^{(k)} > \frac{1}{h}$, the iterative scheme may not converge at all, for, small errors in the current approximations may cause larger errors in the factors $C_{ri}^{(k)}$. It appears that the best we can do is given by the result of the following theorem.

Theorem (7.4): Suppose the conditions of Theorem (7.3) are valid. If $g_{ri}^{(o)}$ is sufficiently close to g_{ri}, $r = 1(1)q$, $i = 1(1)s$, and $C_{ri}^{(k)}$ is chosen so that

$$C_{ri}^{(k)} = \operatorname{sign}\left(\frac{\partial F_r^*}{\partial g_r}\Big|_{(x_i;\, Y_i^{*(k)},\, g_i^{*(k)})}\right) \min\left\{\frac{1}{|h|^{\alpha}},\ \frac{1}{\left|\frac{\partial F_r^*}{\partial g_r}\Big|_{(x_i;Y_i^{*(k)},g_i^{*(k)})}\right|}\right\}$$

$$0 \le \alpha < 1, \quad r = 1(1)q, \quad i = 1(1)s, \quad k = 0,1,\ldots,$$

then there exists $h_o > 0$, such that

$$\lim_{k \to \infty} g_{ri}^{(k)} = g_{ri}$$

for $0 \le h \le h_o$.

Proof: Define

$$\delta_{ri}^{(k)} = g_{ri}^{(k)} - g_{ri}, \qquad r = 1(1)q, \quad i = 1(1)s.$$

Suppose that $(x_i;\, Y_i^{(k)},\, g_i^{(k)})$ lies in N. Then a Taylor expansion for (7.4) gives

$$\delta_{ri}^{(k+1)} = \delta_{ri}^{(k)} - C_{ri}^{(k)} \Bigg[F_r^*\Big(x_i;\ \{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]}\, g_{\rho j}\},\ g_{ri}\Big)$$

$$+ \sum_{\rho=1}^{q} \sum_{m=1}^{n_\rho} \left(\frac{\partial F_r^*}{\partial y_\rho^{(n_\rho - m)}} \Bigg|_{(x_i;\ \underline{y}_i^{(k)},\ \underline{g}_i^{(k)})} \right) \frac{(\mu_i h)^m}{m!} \sum_{j=1}^{s} \lambda_{\rho ij}^{[m]}\, \delta_{\rho j}^{(k)}$$

$$+ \delta_{ri}^{(k)} \left. \frac{\partial F_r^*}{\partial g_r} \right|_{(x_i;\ \underline{y}_i^{*(k)},\ \underline{g}_i^{*(k)})} \Bigg]$$

where $(\underline{x}_i;\ \underline{y}_i^{*(k)},\ \underline{g}_i^{*(k)})$ lies in N.

By (7.2) the first term in square brackets is zero. By setting

$$\delta^{(k)} = \max_{r,i} \left\{ \left| \delta_{ri}^{(k)} \right| \right\},$$

we obtain

$$\left|\delta_{ri}^{(k+1)}\right| \le \left\{ \left| 1 - C_{ri}^{(k)} \left. \frac{\partial F_r^*}{\partial g_r} \right|_{(x_i;\ \underline{y}_i^{*(k)},\ \underline{g}_i^{*(k)})} \right| + K \left| C_{ri}^{(k)}\, h \right| \right\} \delta^{(k)},$$

where K is a bounded positive constant for $(x_i;\ \underline{y}_i^{*(k)},\ \underline{g}_i^{*(k)})$ in N. Since $\frac{\partial F_r^*}{\partial g_r}$, $r = 1(1)q$, are continuous in D, there exists h_1, such that $0 \le h \le h_1$ implies the first term in curlicue brackets is less than some positive number $\frac{\delta}{2}$ $(\delta < 1)$. Further as $C_{ri}^{(k)} < \frac{1}{|h|^\alpha}$, there exists h_2 such that $0 \le h \le h_2$ implies that the second term is less than $\frac{\delta}{2}$. Taking $h_o = \min(h_1, h_2)$ it follows that

$$\delta^{(k+1)} \leq \delta . \delta^{(k)} < \delta^{(k)} .$$

Now if there exists a spherical neighbourhood $N_1 \subseteq N$ of $(x_i; \underline{Y}_i, \underline{g}_i)$ containing the point $(x_i; \underline{Y}_i^{(k)}, \underline{g}_i^{(k)})$, the same is true for the point $(x_i; \underline{Y}_i^{(k+1)}, \underline{g}_i^{(k+1)})$. It follows by induction that the iteration converges to $\underline{g}_i$ if $(x_i; \underline{Y}_i^{(o)}, \underline{g}_i^{(o)})$ is appropriately chosen in N.

Although the points $(x_i; \underline{Y}_i^{*(k)}, \underline{g}_i^{*(k)})$ are not known, $(x_i; \underline{Y}_i^{(k)}, \underline{g}_i^{(k)})$ is usually a sufficiently good approximation that the iteration will still converge to the required solution, and this point will be used in the computation of $C_{ri}^{(k)}$. At the initial point, $\underline{g}_i^{(o)}$ may be taken equal to the vector $\underline{z}_o$ (or the appropriate vector $\underline{g}_o$ in the case that a weight function is used). From a set of values $g_{ri}(x)$ at x already determined, finite difference approximations give for the next step

$$g_{ri}^{(o)}(x+h) = g_{ri}(x) + h\left(\frac{g_{rs}(x) - g_{rl}(x)}{\mu_s h - \mu_l h}\right), \quad r = 1(1)q, \; i = 1(1)s.$$

These appear to be adequate starting approximations.

5. Numerical Examples

The methods described here were applied to several implicit problems. For some of these problems the differential equations could be written explicitly, and corresponding methods for explicit problems were applied. In a comparison of the two types of method for each problem it was found that only a few more iterations were

required to solve the algebraic system in the former case. It appears that the three examples included here may be solved only by appropriate methods for implicit equations. In particular, a singularity occurs in the second example, and none of the approaches described by Collatz [15] appears suitable; however results obtained for this example indicate that the methods described above may be of some use.

For a single weight function satisfying (5.1) and (5.8), it was shown that certain quadrature methods for a class of explicit differential equations are of order 2s-1 or 2s. Here such methods are used, and in two examples the errors are compared with those given by methods based on equally spaced abscissae.

For the examples, we choose $s = 2,3,4$. Values of $g_{ri}^{(o)}$ and $C_{ri}^{(k)}$ are calculated using the formulae of the previous section. A Jacobi-type iteration is used to calculate a new set of iterates using only those of the previous set. Iteration is terminated when the relative error in two sets of iterates is less than 10^{-10}; we remark that substitution of these iterates into the quadratures giving $\hat{y}(x+h)$ further reduces this error for small $|h|$. In the tables the errors are expressed with a base 10 exponent in brackets.

(7.5) $$(y+y')\ln(y+y') + y = 0$$

$$y(0) = 0, \quad y'(0) = 1,$$

has the solution

$$y(t) = te^{-t} .$$

A unit weight function is used with abscissae chosen as zeros of the Legendre polynomial, $P_s(2\mu-1)$. As $\left.\frac{\partial F}{\partial y'}\right|_{t=1} = 0$, $C_{ri}^{(k)}$

becomes very large in the final step. Further as $\frac{\partial F}{\partial y}$ does not have the same type of singularity at this point, different choices of a bound for $C_{ri}^{(k)}$ cause different rates of convergence of the

x	s	Number of Iterations		Error in y
		$C_{ri}^{(k)} \leq \frac{1}{\sqrt{h}}$	$C_{ri}^{(k)} \leq \frac{1}{h}$	
.125	2	10	10	-7.2 (-7)
	3	12	12	4.0 (-10)
	4	12	12	-5.0 (-11)
.500	2	11	11	-2.1 (-6)
	3	10	10	1.7 (-9)
	4	9	9	-5.0 (-11)
1.000	2	116	37	-2.7 (-3)
	3	175	63	-8.6 (-4)
	4	263	96	-3.4 (-4)

TABLE (7.1): Solution of (7.5), $h = .125$ on $[0,1]$.

iterative scheme. For $\alpha = 1$, there is faster convergence than for $\alpha = .5$; if α is not bounded, there is divergence. Further, this singularity causes a significant loss of accuracy in the final step.

$$t^2(y')^5 - y(y')^2 + \frac{y}{512} = 0$$

(7.6)

$$y(0) = 0, \quad \lim_{t \to o} t^{\frac{1}{2}} y'(t) = 2^{\frac{1}{2}},$$

has the solution (Frobenius)

$$y(t) = t^{\frac{1}{2}} (2.82842712475 - 3.4526698 \times 10^{-4} t - 2.0170 \times 10^{-7} t^2$$

$$- 2.36 \times 10^{-10} t^3 - 3.6 \times 10^{-13} t^4 \ldots).$$

		Zeros of $\overline{P}_s(2\mu-1)$		Equally Spaced Abscissae	
x	s	Number of Iterations	Error in y	Number of Iterations	Error in y
.125	2	7	1.1 (-10)	7	8.9 (-10)
	3	7	5.5 (-12)	7	3.6(-12)
	4	7	-1.3 (-11)	7	5.5 (-11)
.500	2	3	1.4 (-10)	3	1.5 (-9)
	3	3	1.1 (-11)	3	3.6 (-12)
	4	3	-1.8 (-11)	3	5.1 (-11)
1.000	2	8	-3.8 (-7)	8	-1.7(-4)
	3	9	-4.4 (-10)	8	-4.3 (-7)
	4	9	-2.1 (-11)	9	-2.4 (-7)

TABLE (7.2): Solution of (7.6), h = .125, on [0,1].

Choosing the weight function $w(t) = t^{-\frac{1}{2}}$, $g_o = 2^{\frac{1}{2}}$, it follows that

$$\frac{\partial F^{\prec}}{\partial g} = 5t^{-\frac{1}{2}} g^4 - 2t^{-1} y g$$

is infinite at $t = 0$; however multiplying the differential equation by $t^{\frac{1}{2}}$, the assumption that $\lim_{t \to 0} t^{-\frac{1}{2}} y(t)$ is finite, gives a finite (non-zero) value to the partial derivative at $t = 0$. Thus the result of Theorem (7.3) is valid.

Here the weight function $\omega(t) \equiv t^{\frac{1}{2}}$ is introduced only to integrate away from the singularity. As the parameters must be recomputed at each step, and further the computation becomes unstable in later steps, $t^{\frac{1}{2}}$ is replaced by a unit weight function beginning at the seventh step. The example is solved by methods based on zeros of the polynomial, $\bar{P}_s(2\mu-1)$, orthogonal with respect to the (appropriate) weight function on $[0,1]$, and by methods based on equally spaced abscissae.

For the first six steps (in which $\omega(t) = t^{\frac{1}{2}}$ is used), there is only a slight difference in the errors. As these are close to the accuracy of the computer, no conclusions may be drawn. However, after the introduction of the unit weight function, the errors are larger for both choices of abscissae. For the method based on zeros of $P_s(2\mu-1)$, the errors are significantly smaller, and thus it appears for a unit weight function this type of method is of higher order.

$$(7.7) \qquad \cos^{-1}\left(\frac{y_1'' + y_1}{2}\right) \sqrt{1 - \left(\frac{y_1'' + ty_2 - t^2}{2}\right)^2} - y_1 = 0$$

$$t(y_2')^2 + y_2'(\sin t - 2y - y_1') + y_1' + y_2 - 2 \sin t = 0$$

$y_1(.5) = .23971\ 27693\ 02$ $\qquad$ $y_2(.5) = .97942\ 55386\ 0$

$y_1'(.5) = .91821\ 68195\ 5$ $\qquad$ $y_2'(.5) = 1.87758\ 25618\ 9$

$y_1''(.5) = 1.51545\ 23544\ 8$

has the solution

$$y_1(t) = t \sin t, \qquad y_2(t) = t + \sin t .$$

A solution is found for this system in the interval $[.5,1]$ using a steplength $h = .0625$. In this interval the factors $c_{ri}^{(k)}$ are bounded by 3 ; this is not true in the interval $[1, 1.5]$ and if a smaller steplength is not used, it is found that the iterative scheme does not converge, for example, when $t = 1.25$. Further, in problems which have oscillatory solutions, there is, on the average, faster variation in the derivatives, and smaller steplengths may be required to obtain comparable accuracy. In the interval chosen there is no change in signs of the required solution components; however in other intervals where a sign change does occur smaller steplengths may be required to get sufficiently efficient iterative schemes.

As for the previous example, the results of Table (7.3) imply that methods constructed using zeros of $P_s(2\mu-1)$ as abscissae are of order higher than those using equally spaced (E.S.) abscissae.

x	s	Number of Iterations		Error in y_1		Error in y_1'		Error in y_2	
		$P_s(2\mu-1)$	E.S.	$P_s(2\mu-1)$	E.S.	$P_s(2\mu-1)$	E.S.	$P_s(2\mu-1)$	E.S.
.5625	2	16	16	1.6(-8)	-2.6(-5)	-1.2(-8)	2.3(-5)	2.7(-8)	6.7(-6)
	3	16	16	9.1(-13)	2.7(-9)	1.8(-11)	-2.4(-8)	3.6(-12)	4.6(-9)
	4	16	16	9.1(-13)	2.4(-9)	1.8(-11)	-1.1(-9)	0	-2.3(-10)
.75	2	10	11	6.7(-8)	-1.1(-4)	-6.2(-8)	9.9(-5)	8.5(-8)	3.3(-5)
	3	11	11	9.1(-12)	4.5(-9)	2.5(-11)	-8.8(-8)	1.5(-11)	4.4(-9)
	4	11	11	7.3(-12)	1.0(-8)	1.5(-11)	-5.6(-9)	3.6(-12)	-1.3(-9)
1.000	2	15	14	1.4(-7)	-2.3(-4)	-1.7(-7)	2.2(-4)	1.2(-7)	7.1(-5)
	3	16	15	3.8(-11)	-5.6(-9)	4.4(-11)	-1.5(-7)	4.4(-11)	-9.0(-9)
	4	17	15	3.1(-11)	2.2(-8)	2.5(-11)	-1.5(-8)	2.2(-11)	-3.6(-9)

TABLE (7.3): Solution of (7.7), h = .0625 on [.5, 1] .

CHAPTER VIII

ERROR CONTROL

Introduction

The choice of method for the numerical solution of a differential problem is often a compromise between accuracy and economy. It is desirable to obtain a solution on some finite interval to a specified tolerance with a minimum of work. Although this contingency is too optimistic, for most well-behaved problems a discrete method with a constant stepsize may be acceptable. For some problems, a large variation of the solution or a derivative occurs within one or more sub-intervals of the domain of integration. Here, certain adjustments to stepsize may lead to a considerable reduction of the work required.

An algorithm is developed to control the magnitude of the accumulation of truncation errors. The arguments used are heuristic. It appears that more precise arguments would lead to a more pessimistic algorithm. The results obtained in the numerical examples exhibit the suitability of the algorithm.

By an m-step strategy for a particular problem we shall mean a choice of method together with a sequence of m steps, possibly of different sizes, which cover the interval. One m-step strategy will be better than another if the maximum value of the accumulated error is smaller for the first than for the second strategy. (Such a criterion could be replaced by the magnitude of the relative accumulated error; however, a result of the next section indicates that this choice may be inferior to that selected.) Further an

m-step strategy will be better than an n-step strategy if $m < n$ and the maximum value of the accumulated error for the first strategy is not greater than that for the second. Except for certain trivial problems, it does not appear possible to determine, a priori, an optimum m-step strategy for large m ($m \geq 10$, for example). This is certainly the case if there exist some self-correcting* strategies for a particular problem. However, it seems reasonable to attempt to develop general techniques which lead to good strategies for at least some classes of problems. Here, any advantage gained from self-correcting strategies is neglected.

The accumulated error at any step is precisely the accumulation of the local (truncation) errors, and may be controlled by appropriately controlling the local errors at each of the previous steps. In general, the local errors depend on current values of the solution and its derivatives as well as the step-size. As these values are not known even approximately until the numerical solution is available, it appears best to control these errors while obtaining the numerical solution. Indeed, this has been the approach of several investigators who have developed criteria for the adjustment of step-size.

However, for single step methods, such techniques (as that developed by Merson [31]) involve extra computation both in terms of arithmetic operations and function evaluations. Nordsieck [34]

---

* A strategy is self-correcting if truncation errors in later steps partially cancel the accumulated errors arising from truncation errors occurring earlier.

has developed a method which is related to a certain multistep method for which no extra function evaluations are required. However, approximations for certain derivatives closely approximate actual derivatives at previous step (or off-step) points, and this may introduce large errors when a rapid variation occurs within several steps. Indeed, it might be expected that for a problem with rapid variation a good method uses function and solution values corresponding to points only in the current integration step.

Here an algorithm for stepsize adjustment is developed for the class of single-step methods proposed earlier. For problems with extreme behaviour, it appears from a numerical example that this approach is quite adequate.

It is convenient to consider only first order systems of differential equations here. This restriction is not necessary, and indeed, the principles may be extended easily to methods for systems of arbitrary orders.

2. Control of Local Error

It has been decided that control of the absolute accumulated error rather than the relative accumulated error may lead to good strategies. This decision is supported by the result of the following theorem.

Theorem (8.1): For linear differential equations, the accumulation of a particular local error is independent of any of the solution values in the interval of integration.

Proof: Using the notation of Chapter II, the local error at $x + kh$ for a first order system is

$$E_r(k) = \hat{Y}_r(x + kh, h) - Y_r(x + kh, h), \quad r = 1(1)q, \; k = 1,2,\ldots .$$

If $B_r(t)$, $r = 1(1)q$, are continuous matrices defining the function evaluations, (2.6) may be written

$$(8.1)\quad \hat{Y}_r(x+kh,h) = A_r^{[1,0]} \hat{Y}_r(x+(k-1)h,h) + hMA_r^{[1,1]} B_r(x+kh)\, \hat{Y}_r(x+kh,h),$$
$$r = 1(1)q, \quad k = 1,2,\ldots ,$$

where the matrices are independent of any of the (approximate) solution values or their derivatives. For explicit methods (8.1) may be written

$$(8.1')\quad \hat{Y}_r(x+kh,h) = A_r\, \hat{Y}_r(x+(k-1)h,h) + hB_{rk}\, \hat{Y}_r(x+(k-1)h,h)$$
$$= \prod_{j=\ell+1}^{k} (A_r + hB_{r_j}) \left[Y_r(x+\ell h,h) + E_r(\ell) \right].$$

Hence the accumulation of the local error $E_r(\ell)$ depends only on the product

$$\prod_{j=\ell+1}^{k} (A_r + hB_{r_j})$$

of matrices which are independent of $Y_r(x+jh, h)$, and the result follows easily.

For a sufficiently small h, a similar result is easily proved for implicit methods. In this case, we have

$$(8.1'') \quad \hat{Y}_r(x+kh,h) = \prod_{j=\ell+1}^{k} ((I - hB_{rj})^{-1} A_r)\left[Y_r(x+\ell h,h) + E_r(\ell)\right] .$$

Thus for linear differential equations the local errors may reasonably be controlled independently* of the solution values. (Indeed, the solution of a problem, for example with oscillations, may require an unnecessarily large number of steps if the relative local error is controlled.) This approach also seems reasonable for mildly non-linear problems, for example, in which the Lipschitz constant is small, or the function evaluation is otherwise insensitive to small changes in the solution.

On the other hand, in a problem such as

$$(8.2) \qquad y' = \frac{1}{y^2 + \delta} \quad , \quad y(-1) = -1$$

for small positive values of δ, local (and accumulated) errors in y near to zero will accumulate relative to the corresponding value of y. For this problem, it would probably be better to control the relative local error in a neighbourhood of $y = 0$. In general, such behaviour may be established from the differential system, and an error control technique suitably constructed if necessary.

For linear differential equations, a further analysis of the accumulation of local errors is possible. Indeed, the theorem

---

* For a problem, for example, with an exponentially increasing solution, the solution may be required with the accumulated error bounded relative to the solution. In this case a relative local error control would probably be better.

implies that the accumulation depends on the eigenvalues of the matrices,

$$(8.3) \qquad A_r + hB_{r_j}, \qquad j = \ell+1(1)k,$$

for explicit methods, and of

$$(8.3') \qquad (I - hB_{r_j})^{-1} A_r, \qquad j = \ell+1(1)k,$$

for implicit methods. If these are <u>all less</u> than unity, the accumulated effect of any local error is one of decay. If at least <u>one</u> eigenvalue is <u>greater</u> than unity the accumulation is a growth of error for at least one component.

In the case of decay, the accumulated error will be dominated by the local errors from points near to the current steo. Here, it appears best to adjust the stepsize so that the local error is (approximately) constant for each step. If a growth of error is expected, the maximum value of the accumulated error is usually at the end of the interval. Here, larger contributions arise from earlier errors. If R is the growth rate per step, a good strategy may be expected if the local errors are bounded so that the k-th bound was R^{k-1} times the first bound. In practice, to obtain R, the maximum eigenvalue of (8.3) or (8.3') is required; as these depend on t in general, such a task is unjustified. Further, for sufficiently large values of m, R^m approximates some power of the natural logarithm base e. Thus we may still expect some improved strategies in this case, if the local errors are all bounded by the same constant.

It is also suggested that non-linear problems be treated in this way, with the provision that the relative local error may be controlled for certain problems as outlined above.

3. Error Estimate

Previously, for single step methods, the local error has been controlled by bounding an estimation of this error. In general, such estimates are not bounds, and it appears that any relative measure of the local error would be suitable. Nordsieck [34] bounds the difference between predicted and subsequent corrected function evaluations. This suggests that a similar estimate may be adopted for single step methods with $\mu_s = \mu_{s+1} = 1$. In particular, we consider the methods developed in Section 4 of Chapter III.

For an s-stage method of order p the local (truncation) error is given by

$$E_r(h) = \hat{y}_r(x+h) - y_r(x+h) = h \sum_{i=1}^{s} \alpha_{ri} \varepsilon_{ri} + \psi_{r\,s+1}$$

where each term on the right side is $O(h^{p+1})$. Also it has been shown that

$$\varepsilon_{ri} = O(h^{\frac{p}{2}+1}), \quad \alpha_{ri} \neq 0, \quad p = 4, 6, 8.$$

Thus the first component of the error is bounded by

$$\bar{E}_r(h) = h \sum_{i=1}^{s} \alpha_{ri} |\varepsilon_{ri}|$$

which, in general, is $O(h^{\frac{p}{2}+2})$. Thus we may expect this term

to exceed $E_r(h)$ in absolute value. Further, a suitable (relative) error estimate may be given by $h\,|\varepsilon_{ri}|$ for some value of i with $\alpha_{ri} \neq 0$. Although ε_{ri} is not known for any value of i, by writing $k_{r\,s+1} = f(x+h;\ \hat{\underline{y}}(x+h))$, it follows that

$$|\varepsilon_{rs}| \equiv |k_{rs} - y_r'(x+h)| = O(h^{\frac{p}{2}+1}) ,$$

and

$$|\varepsilon_{r\,s+1}| \equiv |k_{r\,s+1} - y_r'(x+h)| = O(h^{p+1}) .$$

Thus

$$|\varepsilon_{rs}| = |k_{rs} - k_{r\,s+1}| + O(h^{p+1}) ,$$

and the local error will be controlled by bounding

$$(8.4)\quad e_r(x,h) = h\,|k_{rs} - k_{r\,s+1}| , \qquad r = 1(1)q .$$

As $k_{r\,s+1}(x) = k_{r1}(x+h)$, the only (significant) increase in the computation of this estimate is a subtraction at each step.

In general, $e_r(x,h)$ will be non-zero for the methods considered, and will provide a suitable (relative) error estimate. However, if $f_r(t;\ \underline{y})$ is a function of t alone, or $y_r(t)$ is a polynomial of degree less than $(\frac{p}{2}+1)$, then $e_r(x,h)$ will be identically zero, and the local error is equal to $\psi_{r\,s+1}$. The former case may be adequately treated using high order numerical integration; in the latter case, it is unlikely that step size adjustment will reduce the computation significantly. Such problems are not considered further.

Attempting to estimate an appropriate stepsize at the beginning of each step may involve considerable computation. For most problems, it appears adequate to halve or double the stepsize when appropriate. Thus the stepsize will be halved when $e_r(x,h)$ exceeds some constant, and <u>provisionally</u> doubled when it is less than another. Hopefully, this will ensure that the actual local error is bounded.

The choice of constants for bounding $e_r(x,h)$ is now motivated. First $e_r(x,h)$ may be written

$$(8.5) \qquad e_r(x,h) = \sum_{i=\frac{p}{2}+2}^{p} c_{r\underline{i}}(x,h)h^{i} + O(h^{p+1}) \quad , \quad (\underline{i} = i - \tfrac{p}{2} - 1),$$

where $c_{r\underline{i}}(x,h)$ depend on $x,h,$ and the derivatives and partial derivatives of $\underline{y}$ and $\underline{f}$. Some authors assume that $c_{r1}(x,h)$ is nearly constant, and that for h small $e_r(x,h)$ may be replaced by the first term of this expansion. If difficult behaviour is expected, this assumption will not, in general, be valid. However, for many problems the coefficients $c_{r\underline{i}}(x,h)$ are continuous for a small stepsize (and even have continuous derivatives).

Consider the ratio

$$R_r(x,h) = \frac{e_r(x,2h)}{e_r(x,h)} = 2^{\frac{p}{2}+2}\left\{\frac{c_{r1}(x,2h)+2h\ c_{r2}(x,2h)+\ldots}{c_{r1}(x,h) + h\ c_{r2}(x,h)+ \ldots}\right\}$$

If the computed value of this ratio is approximately $2^{\frac{p}{2}+2}$, then we may reasonably assume that $\left\{c_{r1}(x,h)\cdot h^{p/2+2}\right\}$ is smooth,

and approximates $e_r(x,h)$. Further, if the stepsize is halved when $e_r(x,h)$ exceeds a constant δ_r for some value of r, it may be provisionally doubled when

$$e_r(x,h) \leq \frac{\delta_r}{2^{p/2^{+2}}}, \qquad r = 1(1)q.$$

In this case a reasonable strategy may be expected.

If the ratio differs significantly from $2^{p/2^{+2}}$, then a good strategy may not be expected. Indeed $R_r(x,h) << 2^{p/2^{+2}}$ indicates that $c_{r1}(x,h)$ changes rapidly for small changes in h, and thus that very rapid variation occurs in a neighbourhood of x. (This may possibly be remedied by using a much smaller steplength, and a correspondingly smaller constant to bound $e_r(x,h)$.) If $R_r(x,h) >> 2^{p/2^{+2}}$, probably $e_r(x,h)$ is approximately equal to a later term of expansion (8.5). If $R_r(x,h) \geq 2^{p+1}$, this problem is serious, for in this case $e_r(x,h)$ is not a good estimate of the error $|\varepsilon_{rs}|$. (This difficulty is similar to that in which $e_r(x,h) \equiv 0$, and may be overcome by using a method of higher order.) If $2^{p/2^{+2}} << R_r(x,h) < 2^{p+1}$, adequate results may be possible by reducing the constant bounding $e_r(x,h)$.

4. Algorithm for Stepsize Adjustment

The following algorithm is suggested by the foregoing discussion.

(i) First step:

From control constants δ_r, $r = 1(1)q$, determine the starting stepsize h so that

$$e_r(x,h) \leq \delta_r , \qquad r = 1(1)q.$$

If necessary, this stepsize is reduced to $\bar{h}$ so that

$$R_r(x,\bar{h}) > 2^{p/2+1}$$

and the control constants reset to

$$\delta_r = \min \left\{ \delta_r, \; 2^{p/2+1} \, e_r(x,\bar{h}) \right\} .$$

The ratio $R_r(x,h)$ is obtained to indicate the quality of the strategy (at least at this step).

(ii) Stepsize reduction

For any step at t for which

$$e_r(t,h) > \delta_r, \quad \text{for some } r,$$

the stepsize is halved, and the integration repeated.

(iii) Stepsize increase

For any step at t for which

$$(8.6) \qquad e_r(t,h) \leq \frac{\delta_r}{M_r} , \qquad M_r = 2^{p/2+3} , \qquad r = 1(1)q,$$

the stepsize is provisionally doubled. If, on doubling

$$e_r(t,2h) > \delta_r, \quad \text{for some } r,$$

the stepsize is rejected. Further, we set

$$M_r = \tfrac{1}{2} \left[M_r + R_r(t,h) \right]$$

for k steps (here $k = 5$). Thus, stepsize will be (provisionally)

doubled at later steps only if the doubled stepsize will probably be accepted. Further, in problems involving oscillation, it appears better to require that (8.6) be satisfied at least twice in succession before a provisional doubling occurs. This prevents doubling if $e_r(x,h)$ is small occasionally by chance (possibly if $c_{r1}(x,h)$ is not smooth).

A flow chart indicates the basic form of the program used. If no adjustment of stepsize is required, additional computation consists only of the tests for stepsize adjustment. These in turn require storage of the constants $\{\delta_r\}$ plus a few arithmetic operations.

5. Numerical Examples

Two problems exhibiting different types of difficult behaviour are solved by Nordsieck's method with stepsize adjustment, and that described above. Similar to (8.2) the problem

$$(8.7) \qquad y' = \frac{1}{3y^2+\delta}, \qquad \delta = \frac{1}{512}, \qquad y(-1) = -1,$$

has the solution

$$y(t) = \sqrt[3]{\left(\frac{\delta^3}{27} + \frac{(t-\delta)^2}{4}\right)^{\frac{1}{2}} + \frac{t-\delta}{2}} - \sqrt[3]{\left(\frac{\delta^3}{27} + \frac{(t-\delta)^2}{4}\right)^{\frac{1}{2}} - \frac{t-\delta}{2}}.$$

The method of Table (3.1) is used with three different strategies: absolute local error control, relative local error control, and fixed stepsize, respectively. For each of these and for Nordsieck's method the maximum accumulated error occurred when

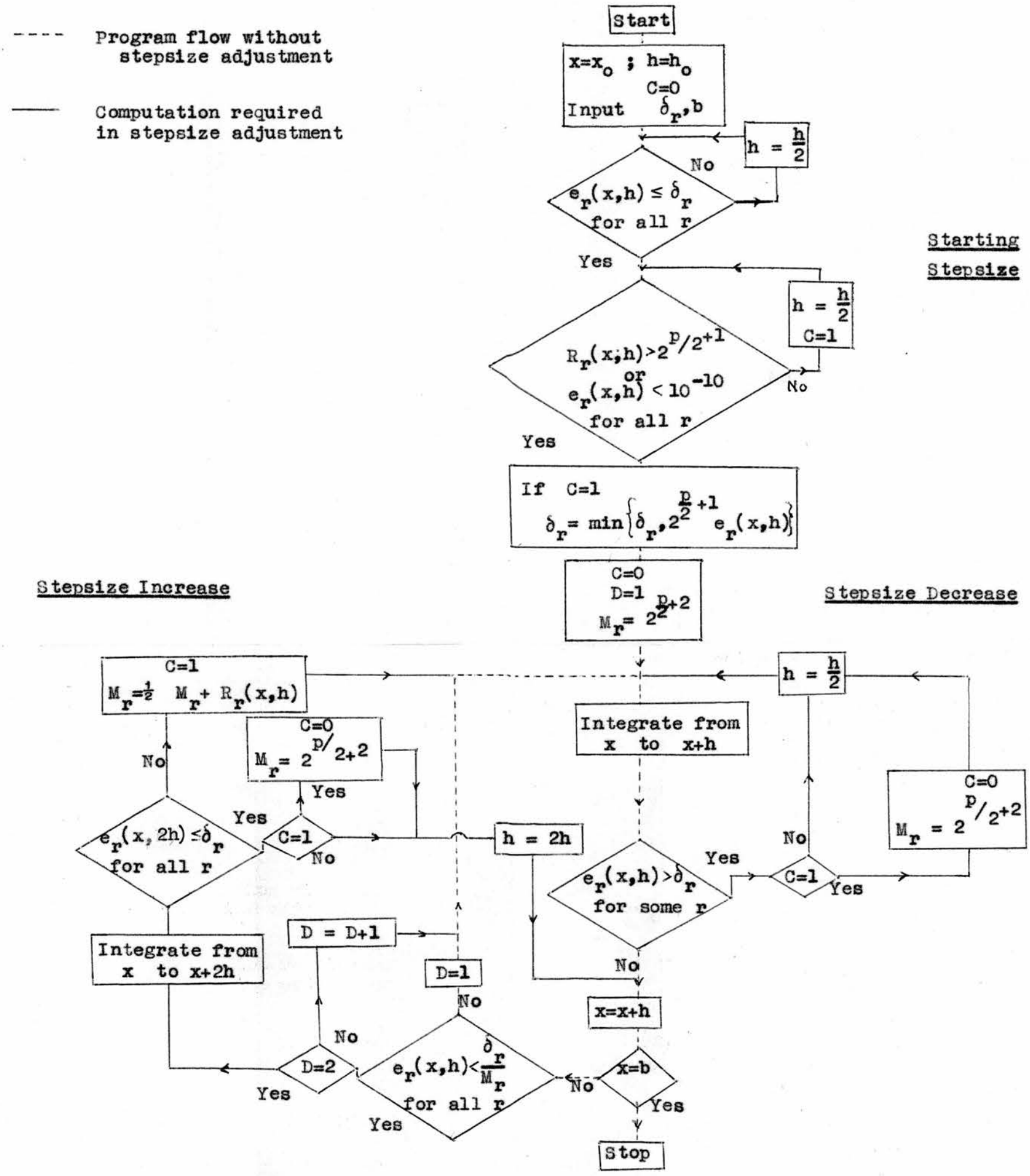

Figure (8.1): Flowchart for adjustment of stepsize.

$\hat{y}(t)$ was a minimum (underlined in Table (8.1)). In the neighbourhood of this value, the accumulated error increased to and decreased from the maximum smoothly.

As expected (slightly) better results are obtained in the neighbourhood of $y(t) = 0$ using a relative local error control than using an absolute error control. However almost twice as many steps are required for the former strategy, and this is not necessarily better than the latter. A fixed stepsize is inadequate. For Nordsieck's method the accumulated errors are slightly larger than using an absolute error control. However fewer than twice as many steps are used, and thus in terms of function evaluations* Nordsieck's method is more economical.

For this problem, the difficult behaviour is restricted to a small neighbourhood of $y(t) = 0$, and although higher derivatives are large in this neighbourhood, they are smooth throughout $(-1, 1)$, and the difficult behaviour can be predicted. In this case, Nordsieck's method which depends significantly on smooth derivatives may be expected to give adequate results. This may not be true if higher derivatives vary significantly in small intervals.

Consider the system

$$y_1' = y_1 - x^5 + 5x^4, \qquad y_1(-1) = -1$$

$$y_2' = 10\pi x^4 \cos(2\pi\, y_1), \qquad y_2(-1) = 0\ , \tag{8.8}$$

---

* Nordsieck's method requires two function evaluations per step. That of Table (3.1) requires seven.

Local Error Control	x	Step Number	h	$\hat{y}$	Error in y
Absolute	-.93750	1	.0625	-.9787313654	2.61(-8)
2.22(-7)	0	56	.000488	-.1197938019	1.00(-6)
	.00195	87	.0000038	.0000242592	2.43(-5)
	.94995	229	.0625	.9816925226	1.55(-8)
Relative	-.93750	1	.0625	-.9787313654	2.61(-8)
2.22(-7)	0	80	.000122	-.1197943027	5.03(-7)
	.00195	171	.00000095	.0000116095	1.16(-5)
	.96050	406	.0625	.9853264752	3.38(-7)
Fixed	-.99130	1	.008696	-.9970949092	3.80(-7)
Stepsize	-.00870	114	.008696	-.2170444790	1.85(-6)
	0	115	.008696	-.1201726107	3.78(-4)
	.00870	116	.008696	-.0808159651	2.66(-1)
	.99130	229	.008696	1.0061828480	1.04(-2)
Nordsieck	-.93750	1	.0625	-.9787313659	2.65(-8)
1.0(-6)	0	104	.000122	-.1197920259	2.78(-6)
	.00195	163	.00000196	.0000645088	6.45(-5)
	.99155	376	.03125	.9958683576	3.32(-7)

TABLE (8.1): Accumulated error for (8.7) on (-1,1).

which has the solution

$$y_1(t) = t^5, \quad y_2(t) = \sin(2\pi t^5) .$$

Here, the difficult behaviour persists throughout the interval, and, indeed, becomes worse as t increases in magnitude. Adequate results could not be obtained using Nordsieck's method on $(-1, 1)$. This may be due to an inadequate starting procedure at $t = -1$ where higher-ordered derivatives vary rapidly.

Much better results are obtained using the single step method with any of the three strategies. The absolute error control gives the best of the strategies. As the Lipschitz constant is large, it could not, a priori, be established that this strategy is better than that for a relative error control. However as all solution values are less then unity, the main difference in the strategies will be a reduction in the stepsize for small solution values for the relative error control. It appears, in general, that a corresponding increase in accuracy is not forthcoming. This suggests that if no direction to the choice of strategy is available, that an absolute error control should be used.

We conclude that if extreme variation occurs in a problem, a single step method is required. Further, to minimize computation for stepsize adjustment, a criterion such as that developed here should be used.

Local Error Control	x	Step Number	h	$\hat{y}_2$	Error in y
Absolute	-.99219	1	.0078125	.2392875551	2.00(-7)
1.0(-5)	-.86719	17	.0078125	-.0601799058	1.18(-6)
	-.02344	64	.0625	.0000000583	1.03(-7)
	.976563	99	.015625	-.6462178822	1.71(-5)
Relative	-.99219	1	.0078125	.2392875557	2.00(-7)
1.0(-5)	-.86719	17	.0078125	-.0601799058	1.18(-6)
	0	126	.00390625	-.0000009704	9.70(-7)
	.99219	237	.015625	-.2393061589	1.88(-5)
Fixed	-.980198	1	.01980198	.5629540687	1.04(-5)
Stepsize	-.861386	7	.01980198	-.1612599071	5.13(-5)
	-.009901	50	.01980198	-.0000423771	4.28(-5)
	.980198	100	.01980198	-.5630318081	8.82(-5)
Nordsieck	-.996094	1	.00390625	.1224142571	9.51(-4)
1.0(-5)	-.867188	33	.00390625	-.5573837602	4.97(-1)
	.003906	112	.015625	19.0025	19.00
	1.000000	217	.00390625	37.7596	37.76

TABLE (8.2): Accumulated error in (8.8) on (-1,1).

CHAPTER IX

ERROR BOUNDS

1. Introduction

If a numerical solution is to provide any information of the actual solution, an error bound is needed. Indeed, it is desirable to obtain a good bound - one which is sharp*, or at least close to the actual error, and economically computed - if possible. Such bounds are not, in general, available. Indeed, one bound for the classical Runge-Kutta method due to Bieberbach [5] requires considerable computation even for one step, and is often pessimistic. Slight improvements to this result are possible (for example, Babuska et alia [4, p. 97] provide a better bound for analytic functions).

It appears that by restricting the class of problems considered, better error bounds may be available. Here, for first order differential equations, we consider the solution on an interval in which the derivatives are monotonic. An easily computed error bound for explicit single step methods is given for problems in which, for example, a Lipschitz constant may be calculated for the whole interval. However, it appears that this bound is still pessimistic.

---

* An error bound is sharp, if it is equal to the error for some problem of the class considered.

2. Error Expressions

Assume for the system

$$y_r' = f_r(x,\underline{y}), \qquad r = 1(1)q, \qquad \underline{y}(x_o) = \underline{y}_o ,$$

the derivatives $y_r'(t)$, $r = 1(1)q$, are monotonic for $t \in [x_o, x_m]$. A set of abscissae $\{\mu_i, i = 1(1)s+1, \mu_1 = 0, \mu_{s+1} = 1\}$ defines points of the interval by

$$x_{ki} = x_k + \mu_i h, \quad x_k = x_o + kh, \quad i = 1(1)s+1, \quad k = 0(1)m.$$

Then $x_{k\,s+1} = x_{k+1} = x_{k+1\,1}$. Approximations $\hat{\underline{y}}(x_{ki})$ to $\underline{y}(x_{ki})$ are defined by

$$\hat{y}_r(x_{ki}) = \hat{y}_r(x_k) + \mu_i h \sum_{j=1}^{i-1} \lambda_{rij}\, \hat{y}_r'(x_{kj}) ,$$

$$\hat{y}_r'(x_{ki}) = f_r(x_{ki}; \hat{\underline{y}}(x_{ki})) .$$

It will be convenient to consider errors* defined by

$$\hat{\psi}_{rki} = y_r(x_k) + \mu_i h \sum_{j=1}^{i-1} \lambda_{rij}\, \hat{y}_r'(x_{kj}) - y_r(x_{ki}) .$$

Here, the accumulated errors are distinguished from corresponding errors at "off-step" points. Hence define

$$\eta_{rk} = \hat{y}_r(x_k) - y_r(x_k) ,$$

$$\varepsilon_{rki} = \hat{y}_r'(x_{ki}) - y_r'(x_{ki}) ,$$

and

---

* By definitions of Chapter III, $\hat{\psi}_{rki} = \psi_{rki} + \mu_i h \sum_{j=1}^{i-1} \lambda_{rij}\, \varepsilon_{rkj}$.

$$\varepsilon_{rki}^{[1]} = \hat{y}_r(x_{ki}) - y_r(x_{ki}) \quad .$$

Then

$$\varepsilon_{rki}^{[1]} = \eta_{rk} + \hat{\psi}_{rki} \quad .$$

Assume that the functions have continuous first partial derivatives, and it follows that

$$(9.1) \quad \varepsilon_{rki} = f_r(x_{ki} ; \hat{\underline{y}}(x_{ki})) - f_r(x_{ki} ; \underline{y}(x_{ki}))$$

$$= \sum_{\rho=1}^{q} \frac{\partial f_r}{\partial y_\rho}(x_{ki} ; \bar{\underline{y}}(x_{ki})) \left\{ \eta_{\rho k} + \hat{\psi}_{\rho ki} \right\} \quad .$$

where the components of $\bar{\underline{y}}(x_{ki})$ lie between those of $\hat{\underline{y}}(x_{ki})$ and $y(x_{ki})$.

3. An Error Bound

We assume that L may be determined, and h is chosen sufficiently small that

$$\left| \mu_i h \sum_{\rho=1}^{q} \frac{\partial f_r}{\partial y_\rho}(x_{ki} ; \bar{\underline{y}}(x_{ki})) \right| \leq L|\mu_i h| < 1 \quad , \quad r = 1(1)q, \quad i = 1(1)s+1 \quad .$$

Then

$$(9.2) \quad |\varepsilon_{rki}| \leq L \cdot \max_\rho \left| \eta_{\rho k} + \hat{\psi}_{\rho ki} \right| \quad .$$

The mean value theorem implies that

$$\hat{\psi}_{rki} = \mu_i h \left(\sum_{j=1}^{i-1} \lambda_{rij} \hat{y}_r'(x_{kj}) - y_r'(\bar{x}_{ki}) \right), \quad i = 1(1)+s+1 \quad ,$$

where

where $\bar{x}_{ki} \in [x_k, x_{ki}]$. As $y_r'(t)$ is monotonic, then $y_r'(\bar{x}_{ki}) \in [y_r'(x_k), y_r'(x_{k\underline{i}})]$ where $\mu_{\underline{i}} \geq \mu_i$, $\underline{i} \leq i$. Then

$$(9.3)\quad \left|\hat{\psi}_{rki}\right| \leq \left|\mu_i h\right| \max_{\ell=1,\underline{i}} \left| \sum_{j=1}^{i-1} \lambda_{rij}\, \hat{y}_r'(x_{kj}) - (\hat{y}_r'(x_{k\ell}) - \varepsilon_{rk\ell}) \right|$$

$$\leq \left|\mu_i h\right| \max_{\ell=1,\underline{i}} \left\{ \left| \sum_{j=1}^{i-1} \lambda_{rij}\, \hat{y}_r(x_{kj}) - \hat{y}_r'(x_{k\ell}) \right| + \left|\varepsilon_{rk\ell}\right| \right\}.$$

For certain methods, a judicious choice of $\underline{i}$ (other than i) may lead to better bounds on $\left|\hat{\psi}_{rki}\right|$. Here, it is convenient to examine the case for $\underline{i} = i$ which is valid for any explicit Runge-Kutta method.

Usually the parameters for this type of method are chosen so that

$$y_r(x_k) + \mu_i h \sum_{j=1}^{i-1} \lambda_{rij}\, y'(x_{kj}) - y_r(x_{ki}) = O(h^{p_i}),$$

where $p_i \geq 1$. Indeed, by writing the solution and derivatives in Taylor series expansions, the left side may be written in terms of higher derivatives. However, bounding of these derivatives is difficult in general, and requires considerable computation when possible. By using (9.3) we hope to minimize the computation required to obtain a bound, even though the bound may be pessimistic.

Substituting (9.3) in (9.2) gives

$$(9.4)\quad \left|\varepsilon_{rki}\right| \leq L \max_{\rho} \left\{ \left|\eta_{\rho k}\right| + \left|\mu_i h\right| \max_{\ell=1,\underline{i}} \left[\left| \sum_{j=1}^{i-1} \lambda_{\rho ij}\, \hat{y}_\rho'(x_{kj}) - \hat{y}_\rho'(x_{k\ell}) \right| + \left|\varepsilon_{\rho k\ell}\right| \right] \right\},$$

$$r = 1(1)q, \quad i = 2(1)s+1 .$$

Also (9.2) implies

$$|\varepsilon_{rk1}| \le L \max_{\rho} |\eta_{\rho k}| = N_k \ .$$

Thus the right side of (9.4) can be maximized over ρ for $\ell=1$. For $\ell=i$, we obtain an implicit inequality. However, the index r on the left side is immaterial, and thus if $\rho = \bar{\rho}$ maximizes the right side with $\ell = i$, it follows that

$$|\varepsilon_{\bar{\rho}ki}| \le (1 - |\mu_i h| L)^{-1} L\left\{|\eta_{\bar{\rho}k}| + |\mu_i h| \left|\sum_{j=1}^{i-1} \lambda_{\bar{\rho}ij} \hat{y}_{\bar{\rho}}(x_{kj}) - \hat{y}_{\bar{\rho}}(x_{ki})\right|\right\}$$

and further, the right side is a bound for $|\varepsilon_{rki}|$, $r = 1(1)q$. Thus (9.4) implies

$$(9.5)\quad |\varepsilon_{rki}| \le E_{ki} = \max_{\rho}\left\{(1+L|\mu_i h|)N_k + L|\mu_i h| \max_{\rho} \left|\sum_{j=1}^{i-1} \lambda_{\rho ij} \hat{y}_{\rho}'(x_{kj}) - \hat{y}_{\rho}'(x_{k1})\right| \right.,$$

$$\left. (1-L|\mu_i h|)^{-1}\left[N_k + L|\mu_i h| \max_{\rho}\left|\sum_{j=1}^{i-1} \lambda_{\rho ij} \hat{y}_{\rho}'(x_{kj}) - \hat{y}_{\rho}'(x_{ki})\right|\right]\right\}.$$

Thus, assuming, for example, that a Lipschitz constant L is available, and that the derivatives are monotonic in $[x_o, x_m]$, (9.5) provides an easily computed bound for $|\varepsilon_{rki}|$. Bounds for the accumulated errors η_{rk+1} are required. Using the bound for $|\varepsilon_{rki}|$ and (9.3), these may be obtained from

$$(9.6)\quad |\eta_{rk+1}| = |\varepsilon^{[1]}_{rk\ s+1}| \le |\eta_{rk}| + |\hat{\psi}_{rk\ s+1}|$$

$$\le |\eta_{rk}| + |h| \max\left\{\left|\sum_{j=1}^{s} a_{rj} \hat{y}_r'(x_{kj}) - \hat{y}_r'(x_{k1})\right| + N_k, \right.$$

$$\left. \left|\sum_{j=1}^{s} a_{rj} \hat{y}_r'(x_{kj}) - \hat{y}_r'(x_{k\ s+1})\right| + E_{k\ s+1}\right\} .$$

Thus only $E_{k\,s+1}$ is required at each step.

Better bounds of a similar nature are possible. For example, by bounding each of the first partial derivatives individually, (9.2) may be rewritten as a matrix inequality. Solution of a related matrix problem yields bounds on ε_{rki}, which are different, in general, for each index r. Further improvement may be possible using $\underline{i}$ different from i in (9.3). However, such improvements require additional computation which does not appear justified.

4. Numerical Example

We consider a single differential equation, and for the classical Runge-Kutta method (9.5) and (9.6) give

$$\eta_o = 0, \qquad E_o^{[1]} = 0,$$

and for $k = 0,1,\ldots,$

$$N_k = L\, E_k^{[1]},$$

$$E_{k\,5} = \max \left\{ (1 + L|h|)N_k + \frac{L\,h}{} \left| \bar{y}'(x_k) - \hat{y}'(x_{k1}) \right| , \right.$$

$$\left. (1 - L|h|)^{-1} \left[N_k + L|h| \left| \bar{y}'(x_k) - \hat{y}(x_{k+1\,1}) \right| \right] \right\} ,$$

$$\left|\eta_{k+1}\right| \le E_{k+1}^{[1]} = E_k^{[1]} + |h| \max \left\{ \left| \bar{y}'(x_k) - \hat{y}'(x_{k1}) \right| + N_k , \right.$$

$$\left. \left| \bar{y}'(x_k) - \hat{y}'(x_{k+1\,1}) \right| + E_{k\,5} \right\} ,$$

where

$$\bar{y}'(x_k) = \frac{1}{6} \left(\hat{y}'(x_{k1}) + 2\hat{y}'(x_{k2}) + 2\hat{y}'(x_{k3}) + \hat{y}'(x_{k4}) \right) .$$

The problem

$$y' = f(t,y) \equiv y - t^5 + 5t^4, \quad y(1) = 1,$$

has the solution

$$y(t) = t^5 .$$

An examination of $f(t,y)$ implies that $y'(t)$ is monotonic in the interval $[1,4]$ at least. Further $L = 1$. If this problem is solved by the classical Runge Kutta formula with $|h| < 1$, the expressions above provide an error bound. For example, for $h = .05$,

$$E_{15} = \underline{.0337} , \qquad |\eta_1| = 1.2 \times 10^{-7} < .0337.$$

Bieberbach's bound 5 for one step is determined by

$$|f(t,y)| \leq N , \quad \left| \frac{\partial^{\ell} f}{\partial t^{i} \partial y^{k}} \right| \leq \frac{M}{N^{k-1}} , \quad (0 < \ell \leq 4) ,$$

$$|\eta_1| \leq E = |h|^5 \left[MN\ 3.680642361 + M^2N\ 5.3618055 + M^3N\ 1.220833 + M^4N\ 0.01666666 \right] .$$

For $h = .05$, using <u>exact</u> maximum values for the partial derivatives in the interval $[1, 1.05]$,

$$N = 5(1.05)^4, \quad M = \frac{60}{5(1.05)^4} ,$$

and

$$E = \underline{.0036} .$$

Neither bound is sharp. Although Bieberbach's bound is better for optimal bounds on the partial derivatives, in general, it is difficult to obtain bounds for these derivatives at all. The bound given here (for a problem with monotonic derivatives) appears more useful as it may be computed easily for all steps concurrently with the numerical solution.

A classification of problems as that above may lead to techniques for obtaining error bounds which are practical. Certainly, to improve the status of numerical solutions for differential equations, considerable research is required in this area.

REFERENCES

1. Abian, A., and A.B. Brown: On the solution of the differential equation $f(x,y,y') = 0$. Am. Math. Monthly 66, 192-199 (1959).

2. - - - - On the solution of simultaneous first order implicit differential equations. Math. Ann. 137, 9-16 (1959).

3. Adams, J.C., and F. Bashforth: An attempt to test the theories of capillary action ... with an explanation of the method of integration employed... . Cambridge University Press 1883.

4. Babůska, I., M. Práger, and E. Vitásek: Numerical processes in differential equations. Czechoslovakia: SNTL translated for Interscience: London, 1966.

5. Bieberbach, L.: On the remainder of Runge-Kutta formulae of the theory of ordinary differential equations. Zeit. angew. Math. Phys. 2, 233-248 (1951).

6. Butcher, J.C.: Coefficients for the study of Runge-Kutta integration processes. J. Aust. Math. Soc. 3, 185-201 (1963).

7. - - - - On Runge-Kutta processes of high order. J. Aust. Math. Soc., 4, 179-194 (1964).

8. - - - - Implicit Runge-Kutta processes. Math. Comp. 18, 50-64 (1964).

9. - - - - Integration processes based on Radau quadrature formulas. Math. Comp. 18, 233-244 (1964).

10. - - - - On attainable order of Runge-Kutta methods. Math. Comp. 19, 408-417 (1965).

11. - - - - A modified multistep method for the numerical integration of ordinary differential equations. JACM 12, 124-135 (1965).

12. - - - - On the convergence of numerical solutions of ordinary differential equations. Math. Comp. 20, 1-10 (1966).

13. - - - - A multistep generalization of Runge-Kutta methods with four or five stages. JACM 14, 84-97 (1967).

14. Cassity, C.R.: Solutions of fifth-order Runge-Kutta equations. SIAM B 3, 598-606 (1966).

REFERENCES (Contd.)

15. Collatz, L.: The numerical treatment of differential equations. Berlin: Springer-Verlag, 1960.

16. Cooper, G.J.: A class of single step methods for systems of non-linear differential equations. Math. Comp. 21, 597-610 (1967).

17. - - - - Interpolation and quadrature methods for ordinary differential equations. Math. Comp. 22, 69-76 (1968).

18. - - - - The numerical solution of stiff differential equations. FEBS Summer School, Edinburgh 1968: Computing techniques in Biochemistry, Vol. 2, Supplement, 22-29 (1969).

19. - - - - Error bounds for some single-step methods. Proc. of Conference on Numerical Solution of Differential Equations, Dundee, 1969.

20. Dahlquist, G.G.: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. 4, 33-53 (1956).

21. Gear, C.W.: Hybrid methods for initial value problems in ordinary differential equations. SIAM B 2, 69-86 (1965).

22. Gragg, W.B. and H.J. Stetter: Generalized multistep predictor-corrector methods. JACM 11, 188-209 (1964).

23. Henrici, P.K. : Discrete variable methods in ordinary differential equations. New York: Wiley, 1962.

24. - - - - Error propagation for difference methods. New York: Wiley, 1963.

25. Hull, T.E., and W.A.J. Luxemburg: Numerical methods and existence theorems for ordinary differential equations. Num. Math. 2, 30-41 (1960).

26. Kolmogorov, A.N., and S.V. Fomin: Elements of the theory of Functions and Functional Analysis, Vol. 1. Translation. New York: Graylock Press, 1957.

27. Kutta, W.: Beitrag zur nächerungsweisen integration totaler differentialgleichungen. Zeit. für Math. Phys. 46, 435-453 (1901).

28. Lambert, J.D., and B. Shaw: On the numerical solution of $y' = f(x,y)$ by a class of formulae based on rational approximation. Math. Comp. 19, 456-462 (1965).

REFERENCES (Contd.)

29. Lawson, J.D.: An order five Runge-Kutta process with extended region of stability, SIAM B 3, 593-597 (1966).

30. - - - - An order six Runge-Kutta process with extended region of stability, SIAM B 4, 620-625 (1967).

31. Merson, R.H.: An operational method for study of integration processes. Proc. of Symposium on Data Processing, Weapons Research Est., S. Aust., Paper 110.

32. Murray, F.J., and K.S. Miller: Existence theorems for ordinary differential equations. New York Univ. Press, 1954.

33. Nemytskii, V.V., and V.V. Stepanov: Qualitative theory of differential equations. Translation. Princeton University Press, 1960.

34. Nordsieck, A.: On numerical integration of ordinary differential equations. Math. Comp. 16, 22-49 (1962).

35. Runge, C.: Uber die numerische auflosung von differentialgleichungen. Math. Ann. 46, 167-178 (1895).

36. - - - - Uber die numerische auflosung totaler differentialgleichungen. Nachr. Ges. Wiss. Göttingen - Math. Phys. Kl., 252-257 (1905).

37. Sarafyan, D.: Existence and approximation theorems for ordinary differential equations and their systems. L'Enseignement Mathematique 11, 137-158 (1965).

38. Shanks, E.B.: Solutions of differential equations by evaluations of functions. Math. Comp. 20, 21-38 (1966).

39. Verner, J.H.: The order of some implicit Runge-Kutta methods. Num. Math. 13, 14-23 (1969).

40. - - - - Implicit methods for implicit differential equations. Proc. of Conference on Numerical Solution of Differential Equations, Dundee, 1969.

41. de Vogeleare, R.: A method for the numerical integration of differential equations of second order without explicit first derivatives. J. Research Nat. Bureau of Standards 54, 119-125 (1955).

APPENDICES

A. Bibliography

B. The order of some implicit Runge-Kutta methods. Reprint from Numerische Mathematik 13 (1969).

C. Implicit methods for implicit differential equations. Copy of an article to appear in the Proceedings of a Conference on the Numerical Solution of Differential Equations at Dundee, Scotland, 1969.

BIBLIOGRAPHY

In addition to the articles cited in the list of references, many treatments of numerical solutions for initial value problems appear in recent literature. A selection of articles published since 1964 are listed together with several less recent, but significant, articles and monographs. The author acknowledges that some of these have contributed to the development of this thesis.

Ansorge, R.: On the structure of certain convergence criteria for the numerical solution of initial value problems. Num. Math. 6, 224-234 (1964).

Axelson, O.: Global integration of differential equations through Lobatto quadrature. BIT 4, 69-86 (1964).

Berezin, I.S., and N.P. Zhidkov: Computing Methods, Vol. 2: Approximate methods of solving ordinary differential equations. Pergamon Press, London, 1965 (272-341).

Brush, D.G., J.J. Kohfeld and G.T. Thompson: Solution of ordinary differential equations using two "off-step" points. JACM 14, 769-784 (1967).

Brown, R.R., J.D. Riley, and M.M. Bennett: Stability properties of Adams-Moulton type methods. Math. Comp. 19, 90-96 (1965).

Bulirsch, R., and J. Stoer: Numerical treatment of ordinary differential equations by extrapolation methods. Num. Math. 8, 1-13 (1966).

- - - - and J. Stoer: Asymptotic upper and lower bounds for results of extrapolation methods. Num. Math. 8, 93-104 (1966).

Byrne, G.D., and R.J. Lambert: Pseudo Runge-Kutta methods involving two points. JACM 13, 114-123 (1966).

- - - - Parameters for pseudo Runge-Kutta methods. CACM 10, 102-104 (1967).

Carr III, J.W.: Error bounds for the Runge-Kutta single step integration process. J. ACM 5, 39-44 (1958).

Cea, J.: Differential equations: the p-implicit discrete approximation method. Rev. Franc. Trait. Inf. 8, 179-194 (1965) (French). (Comp. Review 10540, Sept.-Oct. 1966).

Ceschino, F., and J. Kuntzmann; Numerical solution of initial value problems. Englewood Cliffs, N.J., 1966.

BIBLIOGRAPHY (Contd.)

Chase, P.E.: Stability properties of predictor-corrector methods for ordinary differential equations. JACM 9, 457-468 (1962).

Cooper, G.J., and E. Gal: Single step methods for linear differential equations. Num. Math. 10, 307-315 (1967).

Dahlquist, G.G.: A special stability problem for linear multistep methods. BIT 3, 27-43 (1963).

- - - - A numerical method for some ordinary differential equations with large Lipschitz constants. IFIP, Supplement, Booklet I, 32-36, Edinburgh, 1968.

Day, J.T.: A one-step method for the numerical solution of second order linear ordinary differential equations. Math. Comp. 18, 664-668 (1964).

- - - - A Runge-Kutta method for the numerical integration of the differential equation $y'' = f(x,y)$. Zeit. Angew. Math. Mech. 5, 354-356 (1965).

- - - - A one-step method for the numerical integration of the differential equation $y'' = f(x)y + g(x)$. Comp. J. 7, 314-317 (1965).

- - - - Quadrature methods of arbitrary order for solving linear ordinary differential equations. BIT 6, 181-190 (1966).

Dejon, B.: Stronger than uniform convergence of multi-step difference methods. Num. Math. 8, 29-41 (1966).

- - - - Addendum to 'Stronger than uniform convergence of multi-step difference methods'. Num. Math. 9, 268-270 (1967).

Distefano, G.P.: Causes of instabilities in numerical integration techniques. Int. J. of Comp. Math. 2, 123-142 (1968).

Ehle, B.L.: High order A-stable methods for the numerical solution of differential equations. BIT 8, 276-278 (1968).

Filippi, S., and K. Spicher: Investigations concerning the method of Runge-Kutta-Fehlberg. Electronische Datenverarbeitung 8, 221-229 (1966)(German). (Comp. Review 12429, July-Aug. 1967).

- - - - and W. Glasmacher: On new results for Runge Kutta processes by non-numerical programs. Electronische Datenverarbeitung 10, 16-23 (1968)(German). (Comp. Review 16261, Feb. 1969).

BIBLIOGRAPHY (Contd.)

Fox, L.: The numerical solution of ordinary and partial differential equations. Pergamon Press: London, 1962.

- - - - and D.F. Mayers: Computing Methods for Scientists and Engineers. Oxford Univ. Press: Oxford, 1968 (193-219).

Fyfe, D.J.: Economical evaluation of Runge-Kutta formulae. Math. Comp. 20, 392-398 (1966).

Gates Jr., L.D.: Numerical solution of differential equations by repeated quadratures. SIAM Rev. 6, 134-147 (1964).

Gautschi, W., and H.A. Antosiewicz: Survey of Numerical Analysis, Edited by J. Todd: Numerical methods in ordinary differential equations. McGraw-Hill: London, 1962 (314-346).

Gear, C.W.: Numerical integration of ordinary differential equations Math. Comp. 21, 146-156 (1967).

- - - - The automatic integration of stiff ordinary differential equations. IFIP, Math., Booklet A, 81-85, Edinburgh, 1968.

Glasmacher, W., and D. Sommer: Implicit Runge-Kutta formulas. Westdeutscher Verlag: Koln und Opladen, Germany 1966 (German). (Comp. Review 12828, Sept.-Oct., 1967).

Gorbunov, A.D. and A.H. Tikhonov: Estimates of the error of a Runge-Kutta method and the choice of optimal meshes. USSR C.M.M.P. V.4, No. 2, 30-42 (1964)(Trans.).

- - - - and Yu.A. Shakhov: On the approximate solution of Cauchy's problem for ordinary differential equations to a given number of correct figures, II. USSR C.M.M.P., V.4, No. 3, 37-47 (1964) (Trans.).

Gragg, W.B.: On extrapolation algorithms for ordinary initial value problems. SIAM B 2, 384-403 (1965).

Guerra, S.: High order Runge-Kutta formulas for first order linear differential equations. Calcolo 3, 407-440 (1966)(Italian). (Comp. Review 15234, Sept. 1968).

Gurianov, V.M.: An estimate of the error of the approximate solution of the Cauchy problem for a system of ordinary differential equations. Izv. Vysh. Uchebn. Zavedenu. Mat. 5, 17-22 (1964) (Russian). (Math. Review 2209, Sept., 1965).

- - - - On the connection between the Runge-Kutta method and Picard's method. Prikl. Mat. Meh. 28, 783-786 (1964)(Russian). (Comp. Review 11581, Mar.-April, 1967).

BIBLIOGRAPHY (Contd.)

Hall, G.: The stability of predictor-corrector methods. Comp. J. 9, 410-412 (1967).

Hull, T.E., and R.L. Johnston: Optimum Runge-Kutta methods. Math. Comp. 18, 306-310 (1964).

Ionescu, D.V.; Numerical integration of second-order differential equations. Mathematica (Cluj) 6, 217-232 (1964)(French). (Math. Review 886, Jan., 1967).

Jain, M.K., and K.D. Sharma: Numerical solution of linear differential equations and Volterra's integral equation using Lobatto quadrature formulae. Comp. J. 10, 101-107 (1967).

Karim, A.I.A.: Strongly stable predictor-corrector methods for the solution of systems of differential equations. Electronische Datenverarbeitung 2, 76-83 (1966). (German). (Comp. Review 12841, Sept.-Oct. 1967).

- - - - The stability of the fourth order Runge-Kutta method for the solution of systems of differential equations. CACM 9, 113-116 (1966).

- - - - Criterion for the stability of numerical integration methods for solution of systems of differential equations. Math. and Math. Phys. J. Res. Nat. Bur. Standards, Section B 71B, 91-103 (1967). (Comp. Review 13873, Mar. 1968).

- - - - A theorem for the stability of general predictor-corrector methods for the solution of systems of differential equations. J. ACM 15, 706-711 (1968).

King, R.: Runge-Kutta methods with constrained minimum error bounds. Math. Comp. 20, 386-391 (1966).

Klopfenstein, R.W., and R.L. Crane: A predictor-corrector algorithm with an increased range of absolute stability. J. ACM 12, 227-241 (1965).

- - - - and R.S. Millman: Numerical stability of a one-evaluation predictor-corrector algorithm for numerical solution of ordinary differential equations. Math. Comp. 22, 557-564 (1968).

Kohfeld, J.J., and G.T. Thompson: Multistep methods with modified predictors and correctors. J. ACM 14, 155-166 (1967).

- - - - and G.T. Thompson: A modification of Nordsieck's method using an "off-step" point. J.ACM 15, 390-401 (1968).

BIBLIOGRAPHY (Contd.)

Konen, H.P., and H.A. Luther: Some fifth order classical Runge-Kutta formulas. SIAM Rev. 7, 551-558 (1965).

- - - - and H.A. Luther: Some singular explicit fifth order Runge-Kutta solutions. SIAM B 4, 607-619 (1967).

Krein, S.G., and L.N. Shablitskaya: On the stability of difference schemes for the Cauchy problem. USSR C.M.M.P. V.6, No. 4, 51-73 (1966) (Trans.).

Krogh, F.T.: Predictor-corrector methods of high order with improved stability characteristics. J.ACM 13, 374-385 (1966).

- - - - A test for instability in the numerical solution of differential equations. J.ACM 14, 351-354 (1967).

- - - - A variable step variable order multistep method for the numerical solution of ordinary differential equations. IFIP, Math., Booklet A, 91-95, Edinburgh, 1968.

Lambert, J.D., and B. Shaw: A method for the numerical solution of $y' = f(x,y)$ based on a self-adjusting non-polynomial interpolant. Math. Comp. 20, 11-20 (1966).

- - - - and B. Shaw: A generalization of multistep methods for ordinary differential equations. Num. Math. 8, 250-263 (1966).

Lambert R.J.: An analysis of the numerical stability of predictor-corrector solutions of non-linear ordinary differential equations. SIAM B 4, 597-607 (1967).

Lawson, J.D.: Generalized Runge-Kutta processes for stable systems with large Lipschitz constants. SIAM B 4, 372-380 (1967).

Legras, J.: A numerical method for solving large systems of linear differential equations. Num. Math. 8, 14-28 (1966)(French). (Math. Review, 1974, Feb. 1967).

Lehmann, N.J.: Error bounds for approximate solutions of differential equations. Num. Math. 10, 261-288 (1967)(German). (Comp. Review 15231, Sept., 1968).

Levy, H., and E.A. Baggott: Numerical solutions of differential equations. Dover: New York, 1950.

Lewis, Jr., H.R., and E.J. Stovall Jr.: Comments on a floating point version of Nordsieck's scheme for the numerical integration of differential equations. Math. Comp. 21, 157-161 (1967).

Luther, H.A.: Further explicit sixth order Runge-Kutta methods. SIAM Rev. 8, 374-380 (1966).

BIBLIOGRAPHY (Contd.)

Luther, H.A.: An explicit sixth order Runge-Kutta formula. Math. Comp. 22, 434-436 (1968).

Makinson, G.J.: Stable high order implicit methods for the numerical solution of systems of differential equations. Comp. J. 11, 305-310 (1968).

Martin, W.C., K.C. Paulson and L. Sashkin: A general method of systematic interval computation for numerical integration of initial value problems. C.ACM 9, 754-757 (1966).

Meshaka, P.: Two methods of numerical integration for differential systems. Rev. Franc. Trait. Inf. 7, 135-148 (1964)(French). (Comp. Review 7329, Mar.-Apr., 1965).

Midgley, J.E.: Calculation of subdominant solutions of linear differential equations. SIAM B 3, 56-66 (1966).

Osborne, M.R.: A method of finite difference approximation to ordinary differential equations. Comp. J. 7, 58-65 (1964).

- - - - On Nordsieck's method for the numerical solution of ordinary differential equations. BIT 6, 51-57 (1966).

- - - - A new method for the integration of stiff systems of ordinary differential equations. IFIP, Math., Booklet A, 86-90, Edinburgh 1968.

Pickard, W.A.: Tables for step-by-step integration of ordinary differential equations of the first order. J.ACM 11, 229-233 (1964).

Ralston, A.: Relative stability in the numerical solution of differential equations. SIAM Rev. 7, 114-125 (1965).

Reiner, M.: An integration procedure including error estimation. BIT 5, 164-174 (1965).

- - - - An error estimate for linear difference methods. Num. Math. 7, 277-285 (1965)(German). (Comp. Review 9747, May-June, 1966).

Richards, P.I., W.D. Lanning, and M.D. Torrey: Numerical integration of large, highly-damped, non-linear systems. SIAM Rev. 7, 376-380 (1965).

Rosenbrock, H.H.: Some general implicit processes for the numerical solution of differential equations, Comp. J. 5, 329-330 (1962).

Rosser, J.B.: A Runge-Kutta for all seasons. SIAM Rev. 9, 417-452 (1967).

BIBLIOGRAPHY (Contd.)

Sarafyan, D.: Multistep methods for the numerical solution of ordinary differential equations. Enseignement Math. 12, 69-79 (1966).

Scraton, R.E.: Estimation of the truncation error in Runge-Kutta and allied processes. Comp. J. 7, 246-248 (1964). (See Correspondence, Comp. J. 8, p. 52(1965)).

Shaw, B.: Modified multistep methods based on a non-polynomial interpolant. J.ACM 14, 143-154 (1967).

- - - - Some multistep formulae for special high order ordinary differential equations. Num. Math. 9, 367-378 (1967).

Shil'krut, D.I.: A method for the approximate solution of ordinary differential equations. USSR C.M.M.P., V.5, No. 4, 41-55 (1965)(Trans.).

Shintani, H.: On a one-step method of order 4. J. Sci. Hiroshima Univ. Ser. A-I, Math. 30, 91-107 (1966). (Comp. Review 13249, Nov.-Dec., 1967).

Simon, W.E.: Numerical technique for solution and error estimate for the initial value problem. Math. Comp. 19, 387-393 (1965).

Spijker, M.N.: Convergence and stability of step-by-step methods for the numerical solution of initial value problems. Num. Math. 8, 161-177 (1966).

Stetter, H.J.: Stabilizing predictors for weakly unstable correctors. Math. Comp. 19, 84-89 (1965).

- - - - A study of strong and weak stability in discretization algorithms. SIAM B 2, 265-280 (1965).

Treanor, C.E.: A method for the numerical integration of coupled first order differential equations with greatly different time constants. Math, Comp. 20, 39-45 (1966).

Walston, D.E., and E.R. Waddell: Accelerating convergence of one-step methods for the numerical solution of ordinary differential equations. Inter. J. of Comp. Math. 2, 23-33 (1968).

Waters, J.: Methods of numerical integration applied to a system having trivial function evaluation. C.ACM 9, 293-296 (1966).

Widland, O.B.: A note on unconditionally stable linear multistep methods. BIT 7, 65-70 (1967).

Witty, W.H.: A new method of numerical integration of differential equations. Math. Comp. 18, 497-500 (1964).

BIBLIOGRAPHY (Contd.)

Zonneveld, J.A.: Automatic integration of ordinary differential equations. Stiching Mathematisch Centrum, Amsterdam, Holland, Report R743, May 1963. (Comp. Review 5086, Jan.-Feb., 1964).

IMPLICIT METHODS FOR IMPLICIT DIFFERENTIAL EQUATIONS

J. H. Verner

Department of Computer Science, University of Edinburgh,

Edinburgh, Scotland.

Summary

Quadrature methods are used to obtain numerical solutions of certain systems of implicit differential equations. Several examples indicate the range of application of the methods.

March 1969

1. Introduction

Methods for the numerical solution of a system of differential equations are usually based on the assumption that the equations are respectively explicit in each of the highest-ordered derivatives. For otherwise implicit differential equations, differentiation yields a new system which is linear in the new (highest-ordered) derivatives provided that the new derivatives are sufficiently well-behaved (Collatz [2,p.97]). As the behaviour of derivatives may not easily be determined a priori, it would be prudent to use more direct methods if they were available. Here we present a computationally efficient method for solving implicit differential equations directly. Further a useful reformulation of a basic existence theorem is obtained. The major results stated here may each be proved by application of the implicit function theorem, and are to be found in a more comprehensive treatment elsewhere.

We consider the system of differential equations

$$(1)\quad f_r(t; \underline{y}(t), \underline{z}(t)) = 0, \qquad r = 1\ (1)q,$$

where

$$\underline{y}(t) \equiv \{y^{(n_\rho - m)}(t)\} = (y_1(t),\ldots,y_1^{(n_1-1)}(t);\ldots;y_q(t),\ldots,y_q^{(n_q-1)}(t))$$

is a point of real Euclidean space R_N, $N = \sum_{r=1}^{q} n_r$, and

$$\underline{z}(t) \equiv \{y^{(n_\rho)}(t)\} = (y_1^{(n_1)}(t),\ldots,y_q^{(n_q)}(t))$$

is a point of real Euclidean space R_q. A differential problem is defined by (1) together with initial values $\underline{y}_o = \underline{y}(x)$ and $\underline{z}_o = \underline{z}(x)$ for some value x of the variable t.

2. Existence of Solutions

Problems exhibiting difficult behaviour may be solved numerically be methods employing weight functions. Although the methods are valid whether or not the weight functions are continuous, it may be difficult initially to establish the existence of a unique solution. The following reformulation of Murray and Miller's [5] basic existence theorem provides a partial answer.

Consider a partially open region D of real Euclidean space R_{N+q+1} :

$(t; \underline{y}, F)$ lies in D if $0 \le t < a$,

$$|y_r^{(n_r-m)} - y_{ro}^{(n_r-m)}| < a, \quad r=1(1)q, \ m=1(1)n_r,$$

$$|F_r - F_{ro}| < a, \quad r=1(1)q.$$

Theorem 1: Define $w_r(t)$, $r=1(1)q$, a set of non-negative weight functions which are continuous in the half-open interval $[x,a)$, and a vector $\underline{F}_o$ such that

$$z_{ro} = F_{ro}\, w_r(0), \qquad r=1(1)q.$$

Consider the functions

$$f_r^*(t; \underline{y}, \underline{F}) = f_r(t; \underline{y}, \{w_\rho(t)F_\rho\}), \quad r=1(1)q,$$

in the region D. Suppose that

H1: $f_r^*(t; \underline{y}, \underline{F})$ are continuous in D for $r=1(1)q$.

H2: $\dfrac{\partial f_r^*}{\partial F_\rho}(t; \underline{y}, \underline{F})$ exist and are continuous on D for

$r,\rho = 1(1)q$.

H3: $f_r^*(x; \underline{y}_o, \underline{F}_o) = 0$, and the Jacobian

$$J = \frac{\partial(f_1^*, \ldots, f_q^*)}{\partial(F_1, \ldots, F_q)}$$

is not zero at the point $(x; y_o, \underline{F}_o)$.

H4: $\dfrac{\partial f_r^*}{\partial y_\rho^{(n_\rho-m)}}(t; \underline{y}, \underline{F})$ exist and are continuous on D for

$$r,\rho = 1(1)q, \quad m=1(1)n_\rho.$$

Then there exists a number $b > x$ such that (1) has a solution $\underline{y}(t)$ which has continuous first derivatives in $[x, b]$, and

$$\underline{y}(x) = \underline{y}_o, \qquad \underline{z}(x) = \underline{z}_o;$$

furthermore the solution is unique for each vector $\underline{F}_o$ satisfying the hypotheses.

For discontinuous weight functions, the proof of the theorem implies the existence of continuous functions $\{\phi_r(t;\underline{y}),\ r = 1(1)q\}$ such that (1) is equivalent to the system

$$y_r^{(n_r)}(t) = w_r(t)\,\phi_r(t;\underline{y}(t)), \qquad r = 1(1)q.$$

However, as the weight functions are discontinuous, the question of existence of solutions of this system remains unanswered.

Henceforth we assume (1) satisfies the hypotheses of Theorem 1, and has a unique solution $\underline{y}(t)$ in an interval $I = [x,c]$ such that the derivatives

$$(2) \qquad \frac{d^{p-n_r}}{dt^{p-n_r}} \{w_r^{-1}(t)\, y_r^{(n_r)}(t)\}, \qquad r = 1(1)q$$

exist and are continuous in I.

3. Numerical Methods and Convergence

In a numerical method the differential system is replaced by an algebraic system, the solution of which yields approximations to the highest-ordered derivatives. Relatively small algebraic systems are derived using implicit methods developed by Butcher [1]. Indeed a suitable method is selected from a more general class in which weight functions compensate for difficult behaviour (Cooper [4]).

Let a set of distinct abscissae $\{\mu_i,\ i = 1(1)\ s+1\}$ define points $x_i = x + \mu_i h$ in I. Then defining

$$T_r^{[m]}(\mu h) = \sum_{\tau=0}^{m-1} \frac{(\mu h)^\tau}{\tau!}\, y_{ro}^{(n_r-m+\tau)}, \qquad r = 1(1)q,\ m = 1(1)n_r,$$

a quadrature method gives

$$(3) \quad y_r^{(n_r-m)}(x+h) = T_r^{[m]}(h) + \frac{h^m}{m!} \sum_{i=1}^{s} \alpha_{ri}^{[m]} F_r(x_i) + R_r^{[m]}(h),$$

$$r = 1(1)q,\ m = 1(1)n_r,$$

where

$$F_r(t) = w_r^{-1}(t)\, y_r^{(n_r)}(t), \quad r = 1(1)q,\ m = 1(1)n_r,$$

and the weights $\alpha_{ri}^{[m]}$ may be chosen so that the quadrature error, $R_r^{[m]}(h)$, is of order at least s. Indeed, if the abscissae are chosen to be zeros of the Legendre polynomial $P_s(2\mu-1)$, the error is of order 2s. To find approximations $\hat{\underline{y}}(x+h)$ to (3), the quadrature error is neglected and approximations

$$(4)\quad F_{ri} = F_r(x_i) + \varepsilon_{ri}, \qquad r = 1(1)q,\ \ i = 1(1)s,$$

defined implicitly by

$$(5)\quad f_r\left(x_i;\ \{T_\rho^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}^{[m]} F_{\rho j}\},\ \{w_\rho(x_i)\, F_{\rho i}\}\right) = 0,$$

$$i = 1(1)s\ ,\quad r = 1(1)q.$$

are used in (3).

Theorem 2: Suppose there exists a continuous solution of (1) which is unique to a choice of $\underline{F}_o$, and let a set of weight functions satisfying the hypotheses of Theorem 1 provide continuous derivatives (2). Then there exist continuous functions $F_{ri}(h)$ such that for sufficiently small h

$$F_{ri}(h) = F_r(x_i) + O(h^{s+1}).$$

Thus the approximations (4) defined by (5) are adequate and further convergence of the method follows by a proof similar to that given by Cooper [3].

We restrict attention now to a system in which $y_r^{(n_r)}(t)$ occurs only in the r-th equation, and consider the iteration

$$(6)\quad F_{ri}^{(k+1)} = F_{ri}^{(k)} - C_{ri}^{(k)}\, f_r^*\left(x_i;\{T^{[m]}(\mu_i h) + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\rho ij}^{[m]} F_{\rho j}^{(k)}\},\ F_{ri}^{(k)}\right)$$

$$r = 1(1)q,\quad i = 1(1)s.$$

Theorem 3: Suppose the conditions of Theorem 2 are valid. If $F_{ri}^{(o)}$ is sufficiently close to F_{ri}. $r = 1(1)q$, $i = 1(1)s$, and $C_{ri}^{(k)}$ is chosen so that

$$C_{ri}^{(k)} = \operatorname{sign}\left(\frac{\partial f_r^*}{\partial F_r}\bigg|_{(x_i;\ \underline{y}_i^{*(k)},\underline{F}_i^{*(k)})}\right)\min\left\{\frac{1}{|h|^\alpha},\ \frac{1}{\left|\frac{\partial f_r}{\partial F_r}\right|_{(x_i;\underline{y}_i^{*(k)},\underline{F}_i^{*(k)})}}\right\}$$

$$0 \le \alpha < 1\ ,\ r = 1(1)q,\ \ i = 1(1)s,\qquad k = 0,1,\ldots,$$

where the point of evaluation of the partial derivative is defined by a truncated Taylor series expansion of (6), then there exists $h_o > 0$, such that

$$\lim_{k\to\infty} F_{ri}^{(k)} = F_{ri}$$

for $0 \le h \le h_o$.

In practice, the current iterates are used to evaluate $C_{ri}^{(k)}$, and, for the numerical example, $C_{ri}^{(0)}$ is a sufficiently good approximation. Starting values for the iterates are given adequately by finite difference approximations.

4. A Numerical Example

We consider the problem

$$t^2(y')^5 - y(y')^2 + \frac{y}{512} = 0,$$

$$y(0) = 0, \qquad \lim_{t \to 0} t^{\frac{1}{2}} y'(t) = 2^{\frac{1}{2}}$$

which has the solution (Frobenius)

$$y(t) = t^{\frac{1}{2}}(2.82842712475 - 3.4526698 \times 10^{-4} t - 2.0170 \times 10^{-7} t^2 - 2.36 \times 10^{-10} t^3 - 4.8 \times 10^{-13} t^4 \ldots).$$

ABSCISSAE :		ZEROS OF $P_s(2u-1)$		EQUALLY SPACED	
x	s	no. of iterations	ERROR	no. of iterations	ERROR
.125	2	7	1.1 (-10)	7	8.9 (-10)
	3	7	5.5 (-12)	7	3.6 (-12)
	4	7	-1.3 (-11)	7	5.5 (-11)
.500	2	3	1.4 (-10)	3	1.5 (-9)
	3	3	1.1 (-11)	3	3.6 (-12)
	4	3	-1.8 (-11)	3	5.1 (-11)
1.000	2	8	-3.8 (-7)	8	-1.7 (-4)
	3	9	-4.4 (-10)	8	-4.3 (-7)
	4	9	-2.1 (-11)	9	-2.4 (-7)

Numerical solution of example with $h = .125$ on $[0,1]$.

Using the weight function $w(t) = t^{-\frac{1}{2}}$ and assuming that $\lim_{t \to 0} t^{-\frac{1}{2}} y(t)$ is finite, it may be shown that the result of Theorem 2 is valid. As the use of this weight function requires considerable computation, and further becomes unstable in later steps, a unit weight function is introduced at step seven ; unfortunately this leads to a significant decrease in accuracy for which there appears to be no remedy.

For a unit weight function and abscissae chosen to be zeros of the Legendre polynomial, the author [6] has shown the methods for explicit differential equations are of maximum order. Here, a comparison of errors using zeros of $P_s(2\mu-1)$ and equally spaced abscissae respectively is given in the table (where numbers in brackets represent exponents to base 10).

These results indicate that certain choices of the abscissae lead to more accurate methods for other (than unit) weight functions, even for implicit differential equations.

Acknowledgements: The author is grateful to G. J. Cooper for suggesting this research, and to the National Research Council of Canada for financial support. The author also thanks the Professor of the Department of Computer Science for use of the Department's facilities.

References:

1. Butcher, J. C. : Implicit Runge-Kutta processes. Math.Comp.18, 50-64(1964).
2. Collatz, L. : The numerical treatment of differential equations. Berlin : Springer-Verlag 1960.
3. Cooper, G. J. : A class of single step methods for systems of nonlinear differential equations. Math. Comp. 21, 597-610(1967).
4. --- Interpolation and quadrature methods for ordinary differential equations. Math. Comp.22, 69-76(1968).
5. Murray, F. J., and K. S. Miller : Existence theorems for ordinary differential equations. New York : New York University Press 1954.
6. Verner, J. H. : The order of some implicit Runge-Kutta methods. Num. Math.13, 14-23(1969).

Numer. Math. 13, 14–23 (1969)

The Order of Some Implicit Runge-Kutta Methods

J. H. VERNER

Received February 2, 1968

Abstract. The orders of some single step methods for the solution of a general system of differential equations are established. Leading error terms are given.

1. Introduction

BUTCHER [1] produced methods for the numerical solution of a system of first order differential equations, and showed that some s stage methods have order $2s$. COOPER [3] produced methods for a more general system of differential equations, and showed (with an appropriate definition of order) that a class of s stage methods are of order at least s. For a system of first order equations, a method of the latter class reduces to one of the former; this suggested that for systems of higher ordered equations the methods given by COOPER might be of order greater than s. By assuming the existence of certain ordinary and partial derivatives, for a system of equations whose maximum order is n, it is shown that two types of s stage method have order $2s-n+1$, and that leading error terms can be isolated. The results appear to be of some interest.

Although the algebra is complicated, the principles are simple and we commence with an outline. In Section 2 the methods are described using the notation of COOPER [3] with one slightly different definition. We then obtain implicit equations (2.6) for certain error terms; these equations may be expressed in matrix form (3.1). Provided that a sufficiently small step-length is used, the matrix may be inverted, so that expressions are obtained for the error terms in which all terms of order less than p (the required order of the method) are explicit. Sufficient conditions for a method to be of order p are thereby determined, giving (3.4). In Section 4 the orders of two particular types of method are established. A lemma is first proved using a partial fraction expansion, Eqs. (2.2′), and the fact that a non-singular homogeneous matrix equation has only the zero solution. The results are then established by using a double induction to show that Eqs. (3.4) are satisfied.

Thus we consider the system of ordinary differential equations

$$y_r^{(n_r)}(t) = f_r\left(t; \underline{y}(t)\right), \qquad r = 1(1)\, q, \tag{1.1}$$

where

$$\underline{y}(t) \equiv \{y_\varrho^{(n_\varrho - m)}(t)\} = \left(y_1^{(0)}(t), \ldots, y_1^{(n_1-1)}(t); \ldots; y_q^{(0)}(t), \ldots, y_q^{(n_q-1)}(t)\right).$$

Initial values $\underline{y}(x)$ are given for some value x of the real variable t, and for some given step length h, approximations $\overline{\underline{y}}(x+h)$ to $\underline{y}(x+h)$ are needed.

We shall require a set of s parameters μ_i, $i=1(1)s$, and define $\mu_0=0$, $\mu_{s+1}=1$. For a given step length h, these parameters define a closed interval $[a, b]$,

$$x_i=x+\mu_i h, \qquad i=0(1)s+1,$$
$$a=\min_{i=0(1)s+1}\{x_i\}, \qquad b=\max_{i=0(1)s+1}\{x_i\}.$$

It is assumed, for some positive integer p, that the derivatives $y_r^{(n_r+p)}(t)$, $r=1(1)q$, are continuous for $t\in[a, b]$.

2. Derivation of the Methods

The numerical methods for solving (1.1) are described briefly. Define

$$T_{ri}^{[\nu]}=\sum_{\tau=0}^{\nu-1}\frac{(\mu_i h)^\tau}{\tau!}\,y_r^{(n_r-\nu+\tau)}(x),$$
$$r=1(1)\,q, \qquad \nu=1(1)\,n_r, \qquad i=1(1)\,s+1,$$

and

$$k_{ri}=f_r\left(x_i;\left\{T_{qi}^{[m]}+\frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{qij}^{[m]}\,k_{qj}\right\}\right), \tag{2.1}$$
$$r=1(1)\,q, \qquad i=1(1)\,s.$$

Then an s stage method which provides approximations to $\underline{y}(x+h)$ is defined by

$$\bar{y}_r^{(n_r-\nu)}(x+h)=T_{r,s+1}^{[\nu]}+\frac{h^\nu}{\nu!}\sum_{i=1}^{s}\alpha_{ri}^{[\nu]}\,k_{ri},$$
$$r=1(1)\,q, \qquad \nu=1(1)\,n_r.$$

This defines a class of methods, subclasses of which are defined by constraints placed on parameter values. The subclass of methods considered by Cooper [3] is defined by the constraints

$$\binom{\tau+\nu}{\tau}^{-1}\equiv\frac{\tau!\,\nu!}{(\tau+\nu)!}=\sum_{i=1}^{s}\alpha_{ri}^{[\nu]}\mu_i^\tau, \tag{2.2}$$
$$r=1(1)\,q, \qquad \nu=1(1)\,n_r, \qquad \tau=0(1)\,p'-1,$$

$$\mu_i^\tau\binom{\tau+\nu}{\tau}^{-1}=\sum_{j=1}^{s}\lambda_{rij}^{[\nu]}\,\mu_j^\tau, \tag{2.3}$$
$$i=1(1)\,s, \qquad r=1(1)\,q, \qquad \nu=1(1)\,n_r, \qquad \tau=0(1)\,p''-\nu-1.$$

The order of a method for the numerical solution of a system of differential equations is defined to be p if

$$\bar{y}_r^{(n_r-\nu)}(x+h)-y_r^{(n_r-\nu)}(x+h)=O(h^{p+\nu}), \qquad r=1(1)\,q, \qquad \nu=1(1)\,n_r.$$

It is then shown by Cooper that the order of a method from this subclass is given by p, $p\geqq\min(p', p'')$, and the parameters can be chosen so that $p\geqq s$ (or even $s+1$). Certainly if the μ_i, $i=1(1)s$, are chosen to be distinct points, (2.2) and (2.3) can be satisfied for $\tau=0(1)s-1$; (2.2) for $\tau=0(1)s-1$ gives s

equations in the s unknowns $\alpha_{ri}^{[\nu]}$, $i=1(1)s$, for each value of r and ν; similarly, (2.3) for $\tau=0(1)\,s-1$ gives s equations in the s unknowns $\lambda_{ij}^{[\nu]}$, $j=1(1)s$, for each value of i, r, and ν. For distinct $\{\mu_i\}$, in each case the matrix of coefficients is non-singular, and these equations have unique solutions for $\{\alpha_{ri}^{[\nu]}\}$ and $\{\lambda_{ij}^{[\nu]}\}$ respectively.

We wish to show that for certain choices of the parameters $\{\mu_i\}$, $p>s$, and even $p=2s-n+1$. Indeed, if these parameters are chosen to be zeros of the Legendre polynomial $P_s(2\mu-1)$, BUTCHER [1] has shown for first order systems that $p=2s$. If $\{\mu_i\}$ are chosen in this way, and $\alpha_{ri}^{[\nu]}$ are the weights for the associated Gaussian quadrature on (0, 1), then

$$\sum_{i=1}^{s}\alpha_{ri}^{[1]}\,\mu_i^{\tau}=\int_0^1\mu^{\tau}\,d\mu=\binom{\tau+1}{\tau}^{-1},\qquad \tau=0(1)\,2s-1,$$

and thus for $\nu=1$, (2.2) is valid for $p'=2s$. For his proof, BUTCHER requires further that (2.3) be valid for $\tau=0(1)\,s-1$, and as shown above the parameters $\{\lambda_{ij}^{[1]}\}$ may be chosen so that this is the case. For $\nu>1$, the following lemma shows that there exist solutions of (2.2) with $p'>s$.

Lemma. If the parameters $\{\mu_i\}$, and $\{\alpha_{ri}^{[1]}\}$ are defined so that (2.2) is valid for $\tau<p'$, for $\nu=1$, and

$$(2.2')\quad \alpha_{ri}^{[\nu]}=\frac{\nu}{\nu-1}\,(1-\mu_i)\,\alpha_{ri}^{[\nu-1]},\qquad i=1(1)\,s,\qquad r=1(1)\,q,\qquad \nu=2(1)\,n_r,$$

then (2.2) is valid for $\tau<p'-\nu+1$, for all ν.

Proof. The result is valid for $\nu=1$; we assume (2.2) is valid for $\nu=2(1)m$, and for $\nu=m+1$ we have

$$\begin{aligned}\sum_{i=1}^{s}\alpha_{ri}^{[m+1]}\,\mu_i^{\tau}&=\frac{m+1}{m}\sum_{i=1}^{s}\alpha_{ri}^{[m]}\,(1-\mu_i)\,\mu_i^{\tau}\\&=\frac{m+1}{m}\left[\binom{\tau+m}{\tau}^{-1}-\binom{\tau+m+1}{\tau+1}^{-1}\right]\\&=\binom{\tau+m+1}{\tau}^{-1}\end{aligned}$$

for $\tau+1<p'-m+1$, or $\tau<p'-(m+1)+1$; thus by induction on ν, (2.2) is valid for $\tau<p'-\nu+1$.

By the proof below, a method for which the parameters are chosen as for Gaussian quadratures is shown to be of order $p=2s-n+1$; as COOPER [3] points out this is the maximum attainable order. By not requiring the maximum attainable order, other advantages may be gained. If the $\{\mu_i\}$ are chosen differently, but still distinct, methods of order $2s-n+1-k$, for $k\geqq 1$ may be obtained. For example, in the case of Radau quadrature methods (BUTCHER [2]), which can be extended for systems of differential equations of second and higher orders by (2.2') and (4.1'), one of the μ_i is chosen to be an endpoint of the interval (0, 1), and $k=1$. For methods based on Lobatto quadrature, in which two of the μ_i are chosen to be the two endpoints, $k=2$.

We wish to be able to consider any of these choices, and to do so we find sufficient conditions for a method to be of order p, and in Section 4 continue to use the general indices of (2.2) and (2.3).

We proceed now to obtain explicit expressions for certain errors; in the next section these expressions enable us to establish, in terms of parameter constraints, sufficient conditions for a method to be of order p.

As k_{ri} forms an approximation to $y_r^{(n_r)}(x_i)$, we define errors

$$\varepsilon_{ri} = k_{ri} - y_r^{(n_r)}(x_i), \qquad r = 1(1)\,q, \qquad i = 1(1)\,s,$$

and using (1.1) and (2.1) we obtain

$$\varepsilon_{ri} = \mathscr{f}_r\left(x_i; \left\{T_{\varrho i}^{[m]} + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s} \lambda_{\varrho ij}^{[m]}\left[y_\varrho^{(n_\varrho)}(x_i) + \varepsilon_{\varrho j}\right]\right\}\right) - \mathscr{f}_r\left(x_i; \{y_\varrho^{(n_\varrho - m)}(x_i)\}\right). \tag{2.4}$$

If the expressions in curlicue brackets were approximately equal, ε_{ri} would be small, and thus we define

$$\psi_{\varrho i}^{[m]} = T_{\varrho i}^{[m]} + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s} \lambda_{\varrho ij}^{[m]}\, y_\varrho^{(n_\varrho)}(x_j) - y_\varrho^{(n_\varrho - m)}(x_i),$$

$$\varrho = 1(1)\,q, \qquad m = 1(1)\,n_\varrho, \qquad i = 1(1)\,s.$$

Using Taylor expansions about x, we obtain

$$\psi_{\varrho i}^{[m]} = \sum_{\tau=0}^{p-m-1} \frac{\mu_i^m h^{\tau+m}}{(\tau+m)!}\left[\binom{\tau+m}{\tau}\sum_{j=1}^{s} \lambda_{\varrho ij}^{[m]}\,\mu_j^\tau - \mu_i^\tau\right] y_\varrho^{(n_\varrho+\tau)}(x) + O(h^p). \tag{2.5}$$

We assume from now on that $p \geqq p'' \geqq n = \max_{r=1(1)q}\{n_r\}$ (for otherwise no new results are obtained); then if (2.3) holds we obtain

$$\psi_{\varrho i}^{[m]} = O(h^{p''}).$$

We can now rewrite (2.4) as

$$\varepsilon_{ri} = \mathscr{f}_r\left(x_i; \left\{y_\varrho^{(n_\varrho - m)}(x_i) + \psi_{\varrho i}^{[m]} + \frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s} \lambda_{\varrho ij}^{[m]}\,\varepsilon_{\varrho j}\right\}\right) - \mathscr{f}_r\left(x_i; \{y_\varrho^{(n_\varrho - m)}(x_i)\}\right), \qquad r = 1(1)\,q, \qquad i = 1(1)\,s, \tag{2.6}$$

and if, for example, the functions $\mathscr{f}_r\left(t; z(t)\right)$, $r = 1(1)\,q$, satisfy Lipschitz conditions, Cooper [3] has shown that

$$\varepsilon_{ri} = O(h^{p''});$$

this will be required to establish the order of a method.

3. Order of a Method

We now assume the existence of first and second partial derivatives of $\mathscr{f}_r(t; \{z_\eta^{[\nu]}\})$ with respect to $z_\varrho^{[m]}$, $r = 1(1)\,q$, $\varrho = 1(1)\,q$, $m = 1(1)\,n_\varrho$, and the existence of all derivatives with respect to t of these first partial derivatives. Define

$$\mathscr{f}_r(x_i)_{\varrho m} = \left[\frac{\partial \mathscr{f}_r(x_i; \{z_\eta^{[\nu]}\})}{\partial z_\varrho^{[m]}}\right]_{\{z_\eta^{[\nu]}\} = \{y_\eta^{(n_\eta - \nu)}(x_i)\}}$$

and a Taylor expansion for (2.6) gives

$$\varepsilon_{ri}=\sum_{\varrho=1}^{q}\sum_{m=1}^{n_\varrho}\left[\psi_{\varrho i}^{[m]}+\frac{(\mu_i h)^m}{m!}\sum_{j=1}^{s}\lambda_{\varrho ij}^{[m]}\varepsilon_{\varrho j}\right]\mathscr{f}_r(x_i)_{\varrho m}+\phi_{ri},$$

where

$$\phi_{ri}=O(h^{2p''}).$$

We assume henceforth that $p\leqq 2p''$ (this does not restrict the attainable order).

In matrix notation we can write

$$(I-Z)\,\varepsilon=\theta+\phi, \tag{3.1}$$

where ε is an $s\,q$ column vector, the transpose of which is

$$\varepsilon^T=(\varepsilon_{11},\ldots,\varepsilon_{1s};\ldots;\varepsilon_{q1},\ldots,\varepsilon_{qs}),$$

and also

$$\theta^T=(\theta_{11},\ldots,\theta_{1s};\ldots;\theta_{q1},\ldots,\theta_{qs}),$$

$$\theta_{ri}=\sum_{\varrho=1}^{q}\sum_{m=1}^{n_\varrho}\psi_{\varrho i}^{[m]}\mathscr{f}_r(x_i)_{\varrho m}. \tag{3.1'}$$

The elements of ϕ are of $O(h^{2p''})$, I is the identity matrix, and Z an $sq\times sq$ matrix

$$Z=\{z_{ri,\varrho j}\},$$

where

$$z_{ri,\varrho j}=\sum_{m=1}^{n_\varrho}\frac{(\mu_i h)^m}{m!}\lambda_{\varrho ij}^{[m]}\mathscr{f}_r(x_i)_{\varrho m}.$$

Since $m\geqq 1$, for all sufficiently small h the eigenvalues of Z are of modulus less than unity, and (3.1) gives

$$\varepsilon=(I+Z+Z^2+\cdots+Z^{p-p''-1})\,\theta+O(h^p). \tag{3.1''}$$

For a method to be of order p, the definition requires that

$$\begin{aligned}&\bar{y}_r^{(n_r-\nu)}(x+h)-y_r^{(n_r-\nu)}(x+h)\\&=T_{r,s+1}^{[\nu]}+\frac{h^\nu}{\nu!}\sum_{i=1}^{s}\alpha_{ri}^{[\nu]}\left(\varepsilon_{ri}+y_r^{(n_r)}(x_i)\right)-\sum_{\tau=0}^{p+\nu-1}\frac{h^\tau}{\tau!}\,y_r^{(n_r+\tau-\nu)}(x)+O(h^{p+\nu})\\&=\frac{h^\nu}{\nu!}\sum_{i=1}^{s}\alpha_{ri}^{[\nu]}\varepsilon_{ri}+\sum_{\tau=0}^{p-1}\frac{h^{\tau+\nu}}{(\tau+\nu)!}\left[\binom{\tau+\nu}{\tau}\sum_{i=1}^{s}\alpha_{ri}^{[\nu]}\mu_i^\tau-1\right]y_r^{(n_r+\tau)}(x)+O(h^{p+\nu})\end{aligned} \tag{3.2}$$

be $O(h^{p+\nu})$. If (2.2) holds for $p'\geqq p$, the second term is zero, and it remains only to show that

$$\alpha_r^{[\nu]}\varepsilon=O(h^p),$$

where

$$\alpha_r^{[\nu]}=(0,\ldots,0;\ldots;\alpha_{r1}^{[\nu]},\ldots,\alpha_{rs}^{[\nu]};\ldots;0,\ldots,0).$$

By an induction on l, the general element of Z^{l-1},

$$Z^{l-1}=\{z_{r_1i_1,r_li_l}^{[l-1]}\},\qquad l\geqq 2,$$

is given by

$$z^{[l-1]}_{r_l i_l, r_1 i_1} = \sum_{r_{l-1}=1}^{q} \sum_{i_{l-1}=1}^{s} \cdots \sum_{r_2=1}^{q} \sum_{i_2=1}^{s} z_{r_l i_l, r_{l-1} i_{l-1}} \cdots z_{r_2 i_2, r_1 i_1}. \tag{3.3}$$

Thus the elements of $\alpha^{[\nu]}_{r_l} \varepsilon$ may be obtained from

$$Z^{l-1}\theta = (\delta^{[l-1]}_{11}, \ldots, \delta^{[l-1]}_{1s}; \ldots; \delta^{[l-1]}_{q1}, \ldots, \delta^{[l-1]}_{qs}),$$

$$\delta^{[l-1]}_{r_l i_l} = \sum_{r_1=1}^{q} \sum_{i_1=1}^{s} z^{[l-1]}_{r_l i_l, r_1 i_1} \theta_{r_1 i_1},$$

$$\alpha^{[\nu]}_{r_l} Z^{l-1}\theta = \sum_{i_l=1}^{s} \cdots \sum_{r_1=1}^{q} \sum_{i_1=1}^{s} \alpha^{[\nu]}_{r_l i_l} z_{r_l i_l, r_{l-1} i_{l-1}} \cdots z_{r_2 i_2, r_1 i_1} \theta_{r_1 i_1}.$$

Using a Taylor expansion for $\mathscr{f}_r(x_i)_{\varrho m}$, (3.1′) and (2.5) give

$$\theta_{r\,i} = \sum_{\varrho=1}^{q} \sum_{m=1}^{n_\varrho} \left\{ \sum_{\tau=0}^{p-m-1} \mu_i^m \frac{h^{\tau+m}}{(\tau+m)!} \left[\binom{\tau+m}{\tau} \sum_{j=1}^{s} \lambda^{[m]}_{\varrho i j} \mu_j^\tau - \mu_i^\tau \right] y_\varrho^{(n_\varrho+\tau)}(x) \right\}.$$

$$\left\{ \sum_{\sigma=0}^{p-1-\tau-m} \frac{(\mu_i h)^\sigma}{\sigma!} \mathscr{f}_r^{(\sigma)}(x)_{\varrho m} \right\} + O(h^p)$$

and

$$z_{r\,i,\varrho\,j} = \sum_{m=1}^{n_\varrho} \frac{(\mu_i h)^m}{m!} \lambda^{[m]}_{\varrho i j} \left[\sum_{\sigma=0}^{p-m-1} \frac{(\mu_i h)^\sigma}{\sigma!} \mathscr{f}_r^{(\sigma)}(x)_{\varrho m} \right] + O(h^p)$$

where

$$\mathscr{f}_r^{(\sigma)}(x)_{\varrho m} = \left[\frac{d^\sigma}{dt^\sigma} \mathscr{f}_r(t)_{\varrho m} \right]_{t=x}.$$

To show that $\alpha^{[\nu]}_{r_l} Z^{l-1}\theta$ is of $O(h^p)$, it is therefore sufficient to show that

$$\sum_{i_l=1}^{s} \alpha^{[\nu]}_{r_l i_l} \left[\mu_{i_l}^{m_{l-1}+\sigma_{l-1}} \sum_{i_{l-1}=1}^{s} \lambda^{[m_{l-1}]}_{r_{l-1} i_l i_{l-1}} \left[\mu_{i_{l-1}}^{m_{l-2}+\sigma_{l-2}} \cdots \mu_{i_2}^{m_1+\sigma_1} \sum_{i_1=1}^{s} \lambda^{[m_1]}_{r_1 i_2 i_1} \right.\right.$$
$$\left.\left. \cdot \left[\mu_{i_1}^{m_0+\sigma_0} \left[\binom{\tau+m_0}{\tau} \sum_{i_0=1}^{s} \lambda^{[m_0]}_{r_0 i_1 i_0} \mu_{i_0}^\tau - \mu_{i_1}^\tau \right] \right] \cdots \right] \right] = 0 \tag{3.4}$$

for all possible choices of $\{r_k\} = (r_l, \ldots, r_1, r_0)$, $r_i = 1(1)\,q$,

$$m_k = 1(1)\,n_{r_k}, \qquad \sigma_k = 0(1)\,p - m_k - 1,$$
$$\tau = 0(1)\,\bar{p}, \qquad \bar{p} = p - (m_{l-1} + \cdots + m_0 + \sigma_{l-1} + \cdots + \sigma_0) - 1.$$

If (2.3) holds, the expression in square brackets is zero for $\tau = 0(1)\,p'' - m_0 - 1$. To obtain (3.3) we assumed $l \geqq 2$; however we require (3.4) also for $l = 1$ subject to an appropriate interpretation indicated by the dotted brackets. Thus if (3.4) holds for $l = 1(1)\,p - p''$, $\alpha^{[\nu]}_{r_l} \varepsilon$ has order p, as follows from (3.1″).

4. Methods of Maximum Order

Particular methods of the subclass defined in Section 2 may be defined by adding further constraints to (2.2) and (2.3). For *distinct* parameters μ_i, $i = 1(1)\,s$, we shall show that the two types of method defined by

(i) Eq. (2.2) with $p' \geqq s + n$ for $\nu = 1$ and (2.2′) for $\nu > 1$, and Eq. (2.3) for $\tau = 0(1)\,s - 1$,

and

(ii) Eq. (2.2) with $p' \geqq s+n$ for $v=1$ and (2.2′) for $v>1$, and

$$\frac{\mu_i^\tau}{\tau+1} = \sum_{j=1}^{s} \lambda_{rij}^{[1]} \mu_j^\tau, \tag{4.1}$$

$$i=1(1)\,s, \quad r=1(1)\,q, \quad \tau=0(1)\,s-1,$$

$$\mu_i \lambda_{rij}^{[v]} = \frac{v}{v-1}(\mu_i - \mu_j)\,\lambda_{rij}^{[v-1]}, \tag{4.1′}$$

$$i=1(1)\,s, \quad j=1(1)\,s, \quad r=1(1)\,q, \quad v=2(1)\,n_r,$$

are of order $p'-n+1$ where n is the maximum order of the system of equations. We have to show that parameters satisfying (i) or (ii) also satisfy (3.4). For simplicity we assume that

$$\alpha_{ri}^{[v]} = \alpha_i^{[v]}, \quad \lambda_{rij}^{[v]} = \lambda_{ij}^{[v]}, \quad n_r = n.$$

(In effect this restricts attention to a single differential equation. However, provided (i) or (ii) holds for all r, the proofs given below remain valid for a system of equations.)

Lemma. For parameters satisfying the constraints of (i) or (ii),

$$\sum_{i=1}^{s} \alpha_i^{[v]} \mu_i^{m+\sigma} \left[\binom{m+\tau}{\tau} \sum_{j=1}^{s} \lambda_{ij}^{[m]} \mu_j^\tau - \mu_i^\tau \right] = 0, \tag{4.2}$$

$$m=1(1)\,n, \quad \sigma=0(1)\,p'-v-m, \quad \tau=0(1)\,\bar{p}, \quad \bar{p}=p'-v-m-\sigma.$$

Proof. By (2.2′) we have

$$\sum_{i=1}^{s} \alpha_i^{[v]} \mu_i^{m+\sigma+\tau} = \binom{m+\sigma+\tau+v}{v}^{-1}$$

for $m+\sigma+\tau < p'-v+1$, and we have to show that

$$\sum_{i=1}^{s} \alpha_i^{[v]} \mu_i^{m+\sigma} \sum_{j=1}^{s} \lambda_{ij}^{[m]} \mu_j^\tau = \binom{m+\tau}{\tau}^{-1} \binom{m+\sigma+\tau+v}{v}^{-1} \tag{4.3}$$

for

$$\tau=0(1)\,\bar{p}, \quad \bar{p}=p'-v-m-\sigma.$$

(i) For the first case it is necessary to establish (4.3) only for $\tau=s(1)\,\bar{p}$, $\bar{p}=p'-v-m-\sigma$. As the right side can be written with τ absent from the numerator and terms in the denominator distinct, it can be expanded in partial fractions

$$\frac{m!\,\tau!\,v!\,(m+\sigma+\tau)!}{(m+\tau)!\,(m+\sigma+\tau+v)!} = \sum_{k=1}^{m} \frac{c_k}{\tau+k} + \sum_{l=1}^{v} \frac{d_l}{m+\sigma+\tau+l},$$

the coefficients being independent of τ. Thus

$$\binom{m+\tau}{\tau}^{-1} \binom{m+\sigma+\tau+v}{v}^{-1} = \sum_{k=1}^{m} c_k \sum_{j=1}^{s} \alpha_j^{[1]} \mu_j^{\tau+k-1} + \sum_{l=1}^{v} d_l \sum_{j=1}^{s} \alpha_j^{[1]} \mu_j^{m+\sigma+\tau+l-1} \tag{4.4}$$

holds for $m+\sigma+\tau+\nu-1<p'$. As (4.2) holds for $\tau=0(1)\,s-1$, using (4.3) and (4.4) we obtain

$$\sum_{j=1}^{s}\mu_j^{\tau}\left[\sum_{i=1}^{s}\alpha_i^{[\nu]}\mu_i^{m+\sigma}\lambda_{ij}^{[m]}-\sum_{k=1}^{m}c_k\alpha_j^{[1]}\mu_j^{k-1}-\sum_{l=1}^{\nu}d_l\alpha_j^{[1]}\mu_j^{m+\sigma+l-1}\right]=0,$$
$$\tau=0(1)\,s-1.$$

For distinct parameters μ_j, $j=1(1)\,s$, the matrix of coefficients is non-singular, and hence the expression in square brackets vanishes identically. The value of τ is thus immaterial, and as (4.4) holds for $m+\sigma+\tau+\nu-1<p'$, so also does (4.2).

(ii) For $m=1$, by (2.2) and (4.1) the proof is the same as for (i). Assume that (4.2) holds up to some $m-1\geqq 0$. Then by (4.1′) and (2.2′) we have for m

$$\sum_{i=1}^{s}\alpha_i^{[\nu]}\mu_i^{m+\sigma}\sum_{j=1}^{s}\lambda_{ij}^{[m]}\mu_j^{\tau}=\sum_{i=1}^{s}\alpha_i^{[\nu]}\mu_i^{m+\sigma-1}\sum_{j=1}^{s}\frac{m}{m-1}(\mu_i-\mu_j)\lambda_{ij}^{[m-1]}\mu_j^{\tau}$$
$$=\frac{m}{m-1}\left[\binom{m-1+\tau}{\tau}^{-1}\binom{m-1+\sigma+1+\tau+\nu}{\nu}^{-1}-\binom{m-1+\tau+1}{\tau+1}^{-1}\binom{m-1+\sigma+\tau+1+\nu}{\nu}^{-1}\right]$$
$$=\binom{m+\tau}{\tau}^{-1}\binom{m+\sigma+\tau+\nu}{\nu}^{-1}$$
$$=\binom{m+\tau}{\tau}^{-1}\sum_{i=1}^{s}\alpha_i^{[\nu]}\mu_i^{m+\sigma+\tau},\qquad \tau+m+\sigma+\nu-1<p',$$

(noting that the range of τ is reduced by 1 by both applications of (4.2)) and (4.2) follows by induction on m, and the lemma is proved.

Theorem. For parameters satisfying (i) or (ii), Eqs. (3.4) hold for $p=p_\nu=p'-\nu+1$ giving methods of order $p'-n+1$.

Proof. The lemma establishes (3.4) for $l=1$, and (2.2′) gives

$$\sum_{i_1=1}^{s}\alpha_{i_1}^{[\nu]}\mu_{i_1}^{\tau}=\binom{\tau+\nu}{\nu}^{-1},\qquad \tau+\nu-1<p'.$$

We now assume that (3.4) holds up to some $l-1\geqq 0$, and that

$$\sum_{i_l=1}^{s}\alpha_{i_l}^{[\nu]}\mu_{i_l}^{m_{l-1}+\sigma_{l-1}}\sum_{i_{l-1}=1}^{s}\lambda_{i_li_{l-1}}^{[m_{l-1}]}\mu_{i_{l-1}}^{m_{l-2}+\sigma_{l-2}}\cdots\sum_{i_1=1}^{s}\lambda_{i_2i_1}^{[m_1]}\mu_{i_1}^{\tau}$$
$$=\binom{m_1+\tau}{m_1}^{-1}\cdots\binom{m_{l-1}+\cdots+m_1+\sigma_{l-2}+\cdots+\sigma_1+\tau}{m_{l-1}}^{-1}$$
$$\cdot\binom{m_{l-1}+\cdots+m_1+\sigma_{l-1}+\cdots+\sigma_1+\tau+\nu}{\nu}^{-1},\tag{4.5}$$
$$\tau+(m_{l-1}+\cdots+m_1+\sigma_{l-1}+\cdots+\sigma_1)+\nu-1<p',$$

and proceed to show that corresponding results are valid for l. As well as a result corresponding to (4.5) we have to show that

$$\sum_{i_l=1}^{s}\alpha_{i_l}^{[\nu]}\mu_{i_l}^{m_{l-1}+\sigma_{l-1}}\sum_{i_{l-1}=1}^{s}\lambda_{i_li_{l-1}}^{[m_{l-1}]}\mu_{i_{l-1}}^{m_{l-2}+\sigma_{l-2}}\cdots\sum_{i_1=1}^{s}\lambda_{i_2i_1}^{[m_1]}\mu_{i_1}^{m_0+\sigma_0}\sum_{i_0=1}^{s}\lambda_{i_1i_0}^{[m_0]}\mu_{i_0}^{\tau}$$
$$=\binom{m_0+\tau}{\tau}^{-1}\sum_{i_l=1}^{s}\alpha_{i_l}^{[\nu]}\mu_{i_l}^{m_{l-1}+\sigma_{l-1}}\sum_{i_{l-1}=1}^{s}\lambda_{i_li_{l-1}}^{[m_{l-1}]}\mu_{i_{l-1}}^{m_{l-2}+\sigma_{l-2}}\cdots\sum_{i_1=1}^{s}\lambda_{i_2i_1}^{[m_1]}\mu_{i_1}^{m_0+\sigma_0+\tau},\tag{4.6}$$
$$\tau+(m_{l-1}+\cdots+m_0+\sigma_{l-1}+\cdots+\sigma_0)+\nu-1<p',$$

whence (3.4) holds for l.

Again we must consider the two cases individually.

(i) By replacing τ by $\tau+m_0+\sigma_0$ in (4.5), we obtain for the right hand side of (4.6)

$$\binom{m_0+\tau}{\tau}^{-1}\left[\binom{m_1+m_0+\sigma_0+\tau}{m_1}^{-1}\cdots\binom{m_{l-1}+\cdots+m_0+\sigma_{l-2}+\cdots+\sigma_0+\tau}{m_{l-1}}^{-1}\right. \tag{4.7}$$
$$\left.\cdot\binom{m_{l-1}+\cdots+m_0+\sigma_{l-1}+\cdots+\sigma_0+\tau+\nu}{\nu}^{-1}\right]$$

which has an expansion in partial fractions whose coefficients are independent of τ, and the proof of (4.6) proceeds as in the lemma. To establish (4.5) for l, we increase the subscripts of the indices $\{i_k, m_k, \sigma_k,\ k=0(1)\,l\}$ by 1 in (4.6) with (4.7) replacing the right hand side.

(ii) For $m_0=1$, (4.6) is established as in the proof for (i). Assume (4.6) holds up to some $m_0-1\geqq 0$. For m_0, the right hand side of (4.6) is equal to (4.7) as in (i); by (4.1′) the left hand side is equal to

$$\sum_{i_l=1}^{s}\alpha_{i_l}^{[\nu]}\mu_{i_l}^{m_{l-1}+\sigma_{l-1}}\cdots\mu_{i_2}^{m_1+\sigma_1}\sum_{i_1=1}^{s}\lambda_{i_2 i_1}^{[m_1]}\mu_{i_1}^{m_0+\sigma_0-1}\sum_{i_0=1}^{s}\frac{m_0}{m_0-1}\cdot(\mu_{i_1}-\mu_{i_0})\,\lambda_{i_1 i_0}^{[m_0-1]}\mu_{i_0}^{\tau}$$
$$=\frac{m_0}{m_0-1}\left[\binom{m_0-1+\tau}{\tau}^{-1}\binom{m_1+m_0-1+\sigma_0+1+\tau}{m_1}^{-1}\right.$$
$$\left.-\binom{m_0-1+\tau+1}{\tau+1}^{-1}\binom{m_1+m_0-1+\sigma_0+\tau+1}{m_1}^{-1}\right]\cdot$$
$$\cdot\binom{m_2+m_1+m_0+\sigma_1+\sigma_0+\tau}{m_2}^{-1}\cdots\binom{m_{l-1}+\cdots+m_0+\sigma_{l-1}+\cdots+\sigma_0+\tau+\nu}{\nu}^{-1}$$

which is also equal to (4.7), and (4.6) follows by induction on m_0 for

$$\tau+(m_{l-1}+\cdots+m_0+\sigma_{l-1}+\cdots+\sigma_0)+\nu-1<p'.$$

Again (4.5) is established as in (i).

Thus (3.4) follows in each case by induction on l for $p=p_\nu=p'-\nu+1$, and the order of these two types of method is $p\geqq p'-n+1$.

5. Error Terms

It seems reasonable to investigate the leading error terms to get an estimate of the discretization error. This is possible for those terms of $O(h^{\bar p})$, $p\leqq\bar p<2p''$, as they will be given by (3.2). The term of order $\bar p\geqq p$ in $\alpha_r^{[\nu]}\varepsilon$ is given by

$$E_1(\bar p)=\sum_{l=1}^{\bar p-p''}\sum_{r_0,\ldots,r_{l-1}=1}^{q}\sum_{m_0,\ldots,m_{l-1}=1}^{n_{rj}}\sum_{\sigma_0,\ldots,\sigma_{l-1}=1}^{\bar p-n_{rj}-1}\left\{\sum_{i_l=1}^{s}\alpha_{r i_1}^{[\nu]}\mu_{i_l}^{m_{l-1}+\sigma_{l-1}}\sum_{i_{l-1}=1}^{s}\lambda_{r_{l-1}i_{l-1}}^{[m_{l-1}]}\right.$$
$$\cdots\mu_{i_2}^{m_1+\sigma_1}\sum_{i_1=1}^{s}\lambda_{r_1 i_2 i_1}^{[m_1]}\mu_{i}^{m_0+\sigma_0}\left[\binom{\tau+m_0}{\tau}\sum_{i_0=1}^{s}\lambda_{r_0 i_1 i_0}^{[m_0]}\mu_{i_0}^{\tau}-\mu_{i_1}^{\tau}\right]$$
$$\left.\cdot\frac{y_{r_0}^{(n_{r_0}+\tau)}(x)\,f_{r_1}^{(\sigma_0)}(x)_{r_0 m_0}}{(\tau+m_0)!\,\sigma_0!}\cdot\frac{f_{r_2}^{(\sigma_1)}(x)_{r_1 m_1}}{m_1!\,\sigma_1!}\cdots\frac{f_{r}^{(\sigma_{l-1})}(x)_{r_{l-1}m_{l-1}}}{m_{l-1}!\,\sigma_{l-1}!}\right\},$$
$$\tau=\bar p-(m_{l-1}+\cdots+m_0+\sigma_{l-1}+\cdots+\sigma_0).$$

In addition for $\bar{p} \geqq p' - \nu + 1$, there is a contribution from the second term on the right hand side of (3.2) given by

$$E_2(\bar{p}) = \frac{1}{(\bar{p}+\nu)!}\left[\binom{\bar{p}+\nu}{\nu}\sum_{i=1}^{s} \alpha_{ri}^{[\nu]}\, \mu_i^{\bar{p}} - 1\right] y_r^{(\bar{p}+n_r)}(x).$$

For the leading error terms we obtain

$$\bar{y}_r^{(n_r-\nu)}(x+h) - y_r^{(n_r-\nu)}(x+h) = \sum_{\bar{p}=p'-\nu+1}^{2p''-1} \frac{h^{\bar{p}+\nu}}{\nu!} E_1(\bar{p}) + \sum_{\bar{p}=p'-\nu+1}^{2p''-1} h^{\bar{p}+\nu} E_2(\bar{p}) + O(h^{2p''}).$$

With the possible exception of some simple cases, these expressions are not easily calculated. Estimates of the derivatives may be difficult to obtain, and thus these error terms appear to be of little practical use.

Acknowledgements. The author wishes to thank G. J. COOPER for suggesting this research project and for many constructive comments during the various drafts. The author is also grateful to the National Research Council of Canada for support, and to the Director of the University of London Institute of Computer Science for the use of the Institute's facilities.

References

1. BUTCHER, J. C.: Implicit Runge-Kutta processes. Math. Comp. **18**, 50–64 (1964).
2. — Integration processes based on Radau quadrature formulas. Math. Comp. **18**, 233–244 (1964).
3. COOPER, G. J.: A class of single step methods for systems of nonlinear differential equations. Math. Comp. **21**, 597–610 (1967).
4. — Interpolation and quadrature methods for ordinary differential equations. Math. Comp. **22**, 69–76 (1968).

J. H. VERNER
Department of Computer Science
University of Edinburgh
8 Buccleuch Place
Edinburgh 8, Scotland

Universitätsdruckerei H. Stürtz AG Würzburg
Printed in Germany